REPORT RESUMES

ED 015 224 UD 004 649

GUIDE TO EVALUATION OF TITLE I PROJECTS. DRAFT INFORMATION COPY.

BY- NEIDT, CHARLES O. FRENCH, JOSEPH L.

OFFICE OF EDUCATION (DHEW), WASHINGTON, D.C.

PUB DATE OCT 66

EDRS PRICE MF-$0.50 HC-$4.88 120P.

DESCRIPTORS- *GUIDELINES, *EVALUATION METHODS, *EVALUATION TECHNIQUES, *EDUCATIONAL PROGRAMS, *DATA ANALYSIS, STATISTICAL ANALYSIS, PROGRAM EVALUATION, MEASUREMENT INSTRUMENTS, PROGRAM PROPOSALS, TABLES (DATA), DATA COLLECTION, FEDERAL PROGRAMS, ESEA TITLE I

THESE GUIDELINES ARE FOR THE USE OF LOCAL EDUCATIONAL AGENCIES IN COLLECTING DATA AND FORMULATING DESIGNS TO EVALUATE 1965 ELEMENTARY AND SECONDARY EDUCATION ACT TITLE I PROJECTS FOR DISADVANTAGED PUPILS. PROVISIONS FOR EVALUATION ARE A REQUIRED PART OF EACH PROJECT PROPOSAL, AND EVALUATIVE DATA REPORTED AT THE LOCAL LEVEL ARE SYNTHESIZED AND DISSEMINATED AT STATE AND FEDERAL LEVELS. AS NOTED IN THE GUIDELINES, EVALUATION OF THE EDUCATIONAL ATTAINMENT OF PUPILS PARTICIPATING IN TITLE I ACTIVITIES SHOULD BE IN TERMS OF THE STATED PROGRAM OBJECTIVES AND BEHAVIORAL OUTCOMES, AND THE OBJECTIVES OF A COMPREHENSIVE EVALUATION SHOULD INCLUDE NOT ONLY THE MEASUREMENT OF COGNITIVE ACHIEVEMENTS BUT ALSO OF SOCIAL, EMOTIONAL, AND DEVELOPMENTAL CHANGES. THE EVALUATION DESIGN MAY UTILIZE COMPARISON DATA DERIVED WITHIN THE PROJECT GROUP OR DATA BASED ON VARIOUS FORMS OF EXTERNAL CONTROL GROUPS. THE GUIDELINES DISCUSS DESIRABLE CHARACTERISTICS OF TESTS AND THE USE OF STANDARDIZED TESTS AND SUPPLEMENTARY EVALUATIVE TECHNIQUES. THEY ALSO DESCRIBE SEVERAL PROCEDURES FOR ANALYZING EVALUATION DATA. AMONG THE VARIOUS POSSIBLE PITFALLS IN EVALUATION WHICH ARE DESCRIBED AS HAVING SPECIFIC IMPLICATIONS FOR TITLE I EVALUATIONS IS THE FAILURE TO USE SUFFICIENTLY SENSITIVE EVALUATION INSTRUMENTS. THE GUIDELINES INCLUDE A CONTENT OUTLINE OF AN EXEMPLARY FINAL EVALUATION REPORT, A GLOSSARY OF EVALUATION TERMS, AND A LISTING OF TEST PUBLISHERS AND SELECTED REFERENCES ON MEASUREMENT AND RELATED SUBJECTS. APPENDIXES CONSIST OF PORTIONS OF FIVE EVALUATION SCALES. (LB)

ED015224

DRAFT INFORMATION COPY

Department of Health, Education and Welfare
Office of Education
Washington, D.C. 20202

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

GUIDE TO EVALUATION OF TITLE I PROJECTS

Public Law 89-10
Elementary and Secondary Education Act of 1965

October, 1966

UD 004 649

DISCRIMINATION PROHIBITED -Title VI of the Civil Rights Act of 1964 states: "No person in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subject to discrimination under any program or activity receiving Federal financial assistance." Therefore, Title I of the Elementary and Secondary Education Act of 1965, like every program or activity receiving financial assistance from the Department of Health, Education, and Welfare, must be operated in compliance with this law.

TABLE OF CONTENTS

Page

Author's Preface 1

Evaluation of Educational Projects 2

- Legal Responsibility 3
- Educational Improvement Through Feedback and Generalization 5
- Funding and Annual Reporting 6

The Evaluation Process 8

- Steps in the Process 8
- An Example 9
- Translation of Objectives to Behavioral Outcomes 11
- Content Outline 14
- Process Evaluation 15

Evaluation in Relation to Title I 17

- Importance of Measurement of Change 17
- Standards 18
- Norms 19
- Units 20
- Testing the Culturally Deprived 21

Evaluation Designs 25

- Comparison Data Derived Within the Project Group 26
- Comparison Data Derived Outside the Project Group 29
- Need for Continuous Evaluation Regardless of Comparative Data 32

Content of Evaluation 33

- The Use of Standardized Tests 38
- Supplementary Evaluative Techniques 38
- An Illustrative Taxonomy 42

Sources of Project Data 45

- Division of Groups 48

Preparation of the Proposal 50

Analysis of Data 53

- Raw Scores 53
- Derived Scores 55
- Standard Scores 59
- Standard T Scores 63
- Equality of Units 64
- Descriptive Units 65
- Converting Evaluation Data to Standard Scores 65
- Percentages of Pupils Below a National Percentile 69
- Statistical Significance of Differences 71
- Correlated Samples 73
- Independent Samples 73
- Analysis of Covariance 75
- The Coefficient of Correlation 75
- Final Analysis 80

Pitfalls in Evaluation 81

Preparation of a Final Report 85

Summary Statement 92

Selected References 93

- Reference Volumes 94
- Books 95
- Professional Journals 99
- Test Publishers 100

Glossary 103

Appendix 110

- Geometry Attitude Scale (Method) 111
- English Attitude Scale (Teacher) 112
- Teacher Characteristics Scale 113
- Activities Participation Scale 114
- TV Checklist 115

LIST OF TABLES

Number		Page
1	Computation of Percentiles and Percentile Ranks	58
2	Table of Converting Percentiles to Standard Scores	66
3	Percentages of Pupils Below the 25th and 10th Percentiles on National Norms	70
4	Raw Scores for Twenty Project Pupils on a Pre- and Post-Test of Ability to Apply Principles	74
5	Scores on a Socialization Scale for an Experimental and a Control Group	76
6	Pearson Product Moment Correlations Worksheet for Scores on Two Administrations of a Test	77
7	Spearman Rank Order Correlation worksheet for Scores on Two Administrations of a Test	79

LIST OF FIGURES

Number		Page
1	An example of the "paired comparison" technique for obtaining faculty ratings of an in-service education program.	46
2	Chart illustrating the relationship among the phases of the evaluation of selected portions of hypothetical Title I projects.	51
3	The Normal Distribution.	61
4	Examples of the use of graphs to reflect change.	68
5	Form used in Ohio to summarize the evaluation of project objectives measured by objective data.	87
6	Codes used to complete the form in Figure 5.	88
7	Form used in Ohio for reporting evaluation of project objectives measured by non-standardized instruments.	89
8	Codes used to record the more frequently used objectives for Figure 6.	90

Authors' Preface

This document is designed to provide assistance to educators at the local level in identifying appropriate evaluation designs and in collecting and analyzing appropriate evidence to assess the effectiveness of projects initiated under Title I of Public Law 89-10. Title I clearly states that plans for evaluating the effectiveness of projects are to be an integral part of each proposal submitted. Thus, each local educational agency preparing an application must consider, prior to its initiation, the evaluation procedure to be employed in the proposed project. Where more than one activity or service is included in the project application, the applicant is required to describe his plans to evaluate each activity or service, or set of related activities or services.

Although this document is designed for use by local educational agency personnel, it contains many concepts which will also be applicable at the State level. Since each State educational agency has responsibility for synthesizing the evaluation reports of all projects under its jurisdiction, the early grouping of approved projects according to objectives, designs, and measuring instruments utilized will facilitate greatly the assessment of the overall impact of Title I projects.

Charles O. Neidt

Professor and Head
Department of Psychology and
Director, Human Factors
Research Laboratory
Colorado State University

Joseph L. French

Professor of Special Education
and Educational Psychology
Pennsylvania State University

Evaluation of Educational Projects

Assessing progress toward objectives is of central concern to educators. Assessment of progress (i.e., evaluation) applies to all areas of educational endeavor. Curriculum, instructional methodology, pupil personnel services, public relations, physical plant construction, finance, and research are but a few examples in which evaluation is vital. In its simplest sense, to evaluate is to judge the worth, rate, or value of something. Each decision that is made, each course of action that is chosen, even each word that is spoken follows an evaluation of at least one course of action. Evaluation has taken place anytime something is judged good or bad, better or worse, worth continuing or discontinuing. In education, evaluation provides a basis for making sound decisions about educational practice and procedures. Evaluation is, therefore, a concern of administrators, teachers, specialists, boards of education, state and federal offices, parents, legislators, taxpayer, and all those who carry any responsibility whatsoever for the educational process.

Examples of such decision-making include diagnosing learning difficulties, revising curricular content, granting tenure, and selecting instructional materials. Although there are many possible bases for making decisions, such as custom and tradition, appeal to authority, logic, and personal experience, the concept of collecting evidence on which to base decisions has been by far the most fruitful for educational progress.

The question is not whether or not to evaluate. Rather it is how systematically and objectively the information is to be gathered, how the worth of the evidence is to be judged, and how future activity is to be decided upon. In order to build upon successes and learn from partial successes or failures, we must include the best information we can gather in arriving at decisions. Evaluation of relevant data and finding appropriate relationships among items of information must be handled deliberately and cautiously to determine which objectives are and are not being reached. When objectives are not being reached, the causes should be determined. In some instances we must be patient and anticipate improvement in the future. In other instances, a revision in procedure is in order. By knowing how far one is from a goal and how far he has progressed, it is possible to determine how many more resources to allocate to the task or whether the goal is, in fact, attainable with the procedures employed. As defined here, evaluation involves making judgments with the best evidence available. Sometimes this evidence will be objective test data and other times it will be opinions of worthy judges (such as children in the program or parents or teachers). The evidence should lead to a practical decision. Evaluation necessitates gathering and interpreting evidence to encourage ingenuity and innovation in attaining objectives.

Legal Responsibility

The Congress of the United States recognized the crucial importance of evaluating experimental programs when Title I of the Elementary and Secondary Education Act of 1965 (Public Law 89-10) was approved. Requirements for evaluating educational activities at local, State, and Federal levels of administrative responsibility are stated clearly in the four sections of the Title I legislation which follow:

Section 205 (a) (5)

> that effective procedures, including provision for appropriate objective measurements of educational achievement, will

> be adopted for evaluating at least annually the effectiveness of the programs in meeting the special educational needs of educationally deprived children;
>
> Section 205 (a) (6)
>
> that the local educational agency will make an annual report and such other reports to the State educational agency, in such form and containing such information, as may be reasonably necessary to enable the State educational agency to perform its duties under this title, including information relating to the educational achievement of students participating in programs carried out under this title, and will keep such records and afford such access thereto as the State educational agency may find necessary to assure the correctness and verification of such reports;
>
> Section 206 (a) (3)
>
> that the State educational agency will make to the Commissioner (A) periodic reports (including the results of objective measurements required by section 205 (a) (5)) evaluating the effectiveness of payments under this title and of particular programs assisted under it in improving the educational attainment of educationally deprived children,...
>
> Section 212 (a)
>
> The President shall, within ninety days after the enactment of this title, appoint a National Advisory Council on the Education of Disadvantaged Children for the purpose of reviewing the administration and operation of this title, including its effectiveness in improving the educational attainment of educationally deprived children and making recommendations for the improvement of this title and its administration and operation. These recommendations shall take into consideration experience gained under this and other Federal educational programs for disadvantaged children and, to the extent appropriate, experience gained under other public and private educational programs for disadvantaged children.

The consideration to be given to measurements and evaluation in the preparation of program applications at the local level is further explained in Section 116.22 of "Regulations Applicable to the Administration of Title II of Public Law 81-894 (Title I of Elementary and Secondary Education Act of 1965, P.L. 89-10)" as follows:

Provision for Measurement of Educational Achievement and Evaluation of Programs.

(a) An application by a local educational agency shall describe the procedures and techniques to be utilized in making an evaluation at least annually of the effectiveness of its program under Title II* of the Act in meeting the special educational needs of educationally deprived children, including appropriate objective measurements of educational achievement.

(b) The evaluation of the effectiveness of a program shall include an evaluation of the increase in educational opportunities afforded by such a program as well as by each of the projects comprising that program.

(c) The measurement of educational achievement under such a program shall include the measuring or estimating of educational deprivation of those children who will participate in the program and the comparing, at least annually, of the educational achievement of participating children with some objective standard or norm. The type of measurement used should give particular regard to the requirement on the part of the State that it report to the Commissioner on the effectiveness of the several programs of the participating local educational agencies in the State in improving the educational achievement of educationally deprived children.

(d) The evaluation of programs and projects should, consistent with the nature and extent of participation by children enrolled in private schools, be extended to participation of children enrolled in private schools.

Educational Improvement Through Feedback and Generalization

Title I provides for a complete cycle of educational experimentation and change to take place. In general, this means that, as a first step, local deficiencies will be identified. Next, through carefully evaluated programs, effective procedures for alleviating deficiencies and improving present practices will be developed and demonstrated. Finally, through appropriate dissemination of findings, validated practices will be made available for use in other school systems than those in which the practices were originated.

---

* Title I of Public Law 89-10, legally speaking, amends Public Law 81-874 as Title II of that Act.

Evaluation and reporting are required at four different governmental levels in Title I--local, State, U.S. Office of Education, and a National Advisory Council appointed by the President. At the local level, each educational agency is required to plan for evaluation as an integral part of each project proposed to increase the educational attainment of its disadvantaged pupils. As the experimental project is conducted, it is evaluated locally. Results of local evaluation efforts are reported at least annually to the State level. At the State and Federal levels, synthesis of results and dissemination of findings are required. Any review of Title I at the national level can not be effective unless State and local educational agencies supply the necessary evaluative data. Consequently, it is essential that adequate data be gathered by each local educational agency and that such data be summarized and synthesized by the State educational agency.

The central question involved in the evaluation of Title I projects is: Have the educational attainments of children participating in Title I activities been raised? The only way this question can be answered is to have State and local educational agencies define attainment in behavioral terms so that it is measurable. Thus, the responsibility for adequate evaluation at the local level can not be overemphasized.

Funding and Annual Reporting

Since evaluation is required at least annually, it is essential that baseline or reference data be secured very early in the project period. In some projects, the attainment of specified objectives will not occur in a year or even in several years. Nevertheless, the evaluation of progress toward all objectives should be attempted and reported every year. Descriptions of the increases in educational opportunity provided by Title I (new programs, changes in attitude of teachers, etc.) should also be included in annual reports from local and State educational agencies.

While local educational agencies must assume the responsibility for evaluation, they are not required to supply all of the manpower for evaluation. Specialists in evaluation can be found in many institutions of higher education, in regional educational laboratories, and in State departments of education. The costs of evaluation, including consultant fees (when necessary), can be charged to the Title I project budget. A small investment in evaluation that leads to more effective practices can pay substantial dividends.

The Evaluation Process

Prior to the consideration of evaluation of Title I projects in detail, it is important to recognize the universality of evaluation as a process, independent from the content of any particular educational activity being evaluated. As used here, evaluation is the process of determining the extent to which specified objectives have been reached. Stated in another way, evaluation is the process of assessing the extent and direction of change resulting from an educational experience.

Steps in the Process

The steps in evaluating educational outcomes can be enumerated conveniently as follows:

Step 1. Identification of an educational need in terms of a deficiency, a gap in required competencies, or the absence of some desired behavior.

Step 2. Definition of educational objectives to be achieved through the experience to be evaluated. These objectives should reflect the need which the educational experience is designed to alleviate.

Step 3. Translation of the educational objectives into behavior which will be displayed if the objectives are achieved.

Step 4. Identification of situations in which the presence or absence of the designated behavior can be observed and recorded.

Step 5. Establishment of standards, norms, or units which can be used as interpretive values to reveal absolute or relative amounts of behavior displayed.

Step 6. Selection and consequent application of an evaluation device or devices derived from Steps four and five to all those participating in the educational experience.

Step 7. Analysis of evidence yielded by the evaluation device in terms of progress toward the defined objectives.

Step 8. Statement of conclusions regarding effectiveness in terms of the extent to which objectives were achieved.

Dividing the evaluation process into eight steps is purely arbitrary. The total number of steps is a function of step size and could be three to thirteen, depending on the condensation or expansion of the steps as presented here.

An Example

To illustrate the foregoing steps, let it be assumed that one of the objectives of a swimming class is to be evaluated. Let it be assumed also that all pupils are unable to swim at the start of the class and their inability to swim constitutes the "educational need" referred to in Step 1.

Objective: To teach pupils enrolled in the class to swim.

Translation to behavior: Students who have reached this objective will be able to swim under indoor pool conditions.

Situation: After an instructional period, each pupil will be given a chance to swim in an enclosed and uncrowded pool. No diving will be involved.

Standard: Each pupil must swim 25 yards using any stroke he chooses without touching the bottom of the pool.

Application: Each pupil attempts to swim 25 yards.

Analysis: The number of pupils reaching or exceeding the standard is recorded.

Conclusion: Based on the results of the analysis, a generalization regarding the effectiveness of the instruction is made.

It should be noted that in the foregoing example, only one objective was evaluated. Other objectives that may have been defined for the course, such as knowledge of water safety or stimulating interest in swimming, etc., would require separate evaluation. Also, the assumption that all pupils were unable to swim at the start of the class made it possible to employ one measurement and an absolute standard in the evaluation. In most educational situations, such an assumption is unrealistic. Therefore, it is necessary to measure pupil behavior at the start of the educational experience as well as at the end of the experience. In this manner, relative progress can be assessed.

In one community, Title I funds are being used to give breakfast to approximately 100 disadvantaged children and to provide partial help toward clothing the youngsters for winter. Eighty children, most of whom have never had their teeth checked, received dental services. School personnel identified 20 children with reading difficulties who required and received treatment for poor eyesight. While the obvious but global objective is to improve educational attainment, other objectives in this program involve improved attendance and attitude so that more learning can take place. The translation in Step three should include several things in addition to improved reading skill; such as "students will attend regularly and like school more," "students will not deface the building and materials," "more parents will attend PTA meetings and participate in more school activities." In addition to measuring reading skill, the translations suggest comparing average daily attendance records of this year with an average for the past three years, pre and post measures of attitude, comparing building and material maintenance costs of this year with a previous period and/or another school, recording and comparing student fights and disciplinary action of this year with previous years and/or other schools, recording the trend of parent participation in group meetings and individual conferences, comparing comments by parents and teachers on report cards for this sample with student report cards in the past and with other samples.

Designs for the comparisons suggested above will be presented in a later section. Additional examples are suggested in a different form in Figure 2 on page 51.

Translation of Objectives to Behavioral Outcomes

Before going into further detail about the total evaluation process, it is desirable to examine more closely the key to that process, i.e., an effective statement of objectives. A project without clearly stated objectives is like the proverbial ship without a rudder.

Objectives for a project grow out of observed needs. Sometimes it is barely possible to recognize that a need exists; at other times, a need can be sensed but is vague and undefined; and at still other times, many needs are readily apparent. Some thought and discussion are usually necessary to understand fully and to clarify the perceived inadequacy, gap, problem, or need. Needs can be determined from "felt needs" expressed by pupils; from discrepancies in performance when some group is compared with other pupils at the same level; from interferences with learning as noted by teachers or other observers; or demands on "graduates" of educational programs at any level made by teachers at higher grade levels, by employers or by other members of society such as community leaders or other responsible citizens.

Hopefully, any given need will become well recognized and defined as those closely associated with the pupils and teachers concerned discuss a current situation. As focusing continues, an assessment of resources should take place. Administrators and personnel from related fields can be helpful as discussions become more broadly based.

When the shortcomings which were vaguely described initially can be identified as educational needs, general project objectives can be stated (Step 2). As educational needs dictate, the general objectives can take the form of several relatively independent statements. Although "To employ personnel to provide classes in remedial reading and speech therapy and guidance services" combines several general objectives, it is an undesirable statement because it suggests that the desired end product is to employ teachers or specialists rather than to help children. It would be far more desirable to refer to behavior which can be observed in specific children, such as "To improve the reading and language skills and attitudes toward school of elementary age children."

The major purposes of stating objectives are to insure 1) that the activities planned will be designed to reduce the real need and 2) that the selected means of evaluation will have a direct relationship to the project.

Statements of objectives should be concise and to the point. Whereas the statement of general objectives will usually be a listing of expected academic achievement or anticipated changes in attitude, interest, habits, or adjustment, it is the preliminary step to stating objectives specifically in behavioral outcomes.

In Step three, the objectives become more specific and should be stated in terms of behavior that can be observed. At this point each objective should be translated into at least one learning outcome. Most learning outcomes develop sequentially. For the statement of general objective, "to improve auditory discrimination and reading skills among children in their initial year of school," the translation may read "the children will recognize the sound of words from preprimer and primer readers and respond with words that rhyme with them." "The children will read each word on the Dolch list and pronounce it correctly." In addition, during consecutive phases of the instruction period, the extent to which children can recognize printed words and can name words that rhyme with them can be observed.

In addition to the translations given, other translations from the above objective could be generated. A partial list might include "recognizing and pronouncing consonants, vowels, and blends; developing left to right eye movements; arousing and sustaining interest in reading; and developing word attack skills." Each outcome may have its own pattern of development, but each outcome need not depend upon another outcome. However, all major objectives of each project should be stated. Such a procedure greatly facilitates the assessment process.

An essential ingredient of a valid assessment is a representative sampling of behavior that indicates the degree of achievement in a particular situation. Here, the word "representative" has two meanings: one pertaining to representation of all anticipated outcomes or objectives, and the other pertaining to an adequate sampling of behavior throughout an appropriate range of the characteristic concerned.

An example of evaluating more than the basic objective was given on pages 12 and 13 in relation to improving reading and attitude toward school through physical conditions. The portions of method, teacher, teacher characteristics, and activities participation attitude scales found in the Appendix provide an example of measurement throughout a broad range of concern. The appropriate range should be defined in a "content outline."

Content Outline

The preparation of a complete outline, including a general statement of objectives (Step 2) and their translation into behavioral terms (Step 3), is a highly desirable procedure for relating educational practice to evaluation. The statements in behavioral terms will usually require enumeration or subclassifications. Merely "to improve attitude toward school" is not sufficiently clear. Additional statements pertaining to teachers, buildings, children, courses, activities, and/or materials might be described in behavioral terms, depending upon the real educational need. The process of outlining the objectives is known as the formulation of a "content outline."

The content outline should include each aspect of the perceived need and should suggest its relative importance in relation to other objectives. By stating the need of students in considerable detail and in terms of anticipated behavior, activities can then be selected which will fulfill the need. Activities derived from the specific objective suggest the situations necessary for observing evidence required in Step four. Considerations of evaluation in this document will be primarily concerned with end products, but the content outline will be of great help to teachers as they direct

Title I activities. (The mechanics of organizing statements of need, objectives, behavioral outcomes, and suggested evaluation procedures will be described in the section of this document entitled "Preparation of the Proposal.")

Unless the objectives are fully understood by teachers, project efforts will shift direction with each new wind current. In such situations it is difficult to make much headway or to evaluate progress. Hours of such travel do not indicate progress toward a destination.

Process Evaluation

As pupils participate in the suggested activities, teachers observe efforts of the pupils and evaluate them. This type of evaluation is sometimes referred to as process evaluation.

In this evaluative effort neither the teacher nor the evaluator exercises experimental control over the situation. Instead, they focus attention on aspects of the project that are crucial to its success. In the role of process evaluator one does not intervene in activities or manipulate personnel. However, observations are made daily or weekly, systematically organized, and reported to the director as often as necessary. A summary of these reports, in log form, at the end of a project can be very helpful in deciding when modification of procedure should be employed in future projects. It may be noted that after a field trip the students exhibited much more interest in class discussions. Next year the field trip may be scheduled earlier.

Review of observational records during a project will sometimes suggest modifications in procedure to overcome unanticipated events and keep the project moving toward its objectives. The evaluator may note that teacher-made test scores were low and suggest teaching the concepts again with different techniques before going on with the planned activities. The more lucidly the objectives are stated, the easier will be continuous or periodic sub-evaluation. As teachers continously observe the effect of their work, they can change or modify the procedures to accomplish more effectively the basic objectives. When the objectives are indistinct, entangled, and confused, activities tend to be random and time-filling, rather than directed toward accomplishing the desired end product.

It may be very helpful if pupils understand the objectives and accept them. Such knowledge helps them realize the purpose of the activity and helps direct their behavior toward the end product.

Evaluation in Relation to Title I

In planning the evaluation of Title I projects, three basic considerations are appropriate: first, the development of the design or evaluation procedure for assessing the extent to which objectives have been achieved; second, the identification of the content of measuring instruments or evaluation devices to be utilized; and third, the designation of the source of the evaluation data to be obtained.

For each of these considerations, there are almost unlimited possibilities from which selections for evaluating any specific educational experience can be made. The purpose here is not to provide exhaustive lists or classifications of possible choices, but rather, to describe and illustrate several major categories of designs, measuring devices, and data sources which will be helpful in evaluating experimental educational experiences.

It is anticipated that in some instances evaluators may develop more refined and sophisticated designs and sources of data than those included here. To the extent that such procedures and instruments provide more valid evidence of effectiveness than the types of design, devices, and data sources mentioned here, these procedures and devices should be employed. It is also anticipated, however, that the designs, devices, and data sources mentioned in the following sections will encompass the great majority of evaluation situations encountered under Title I.

Importance of Measurement of Change

Under the provisions of Title I, elementary and secondary schools are able to strengthen and improve the educational opportunities of children with "special educational needs." The majority of objectives of proposed projects will relate to characteristics which are already present to some degree in the pupils studied prior to the initiation of the new educational experiences. It will be readily apparent, therefore, that most of the evaluation procedures and designs appropriate for Title I involve assessing the amount of change in pupil behavior over time. This means, in most instances, that evaluation procedures will involve obtaining not only a baseline or initial measure at the start of the experimental experience, but also one or more subsequent or progress measures as the experience proceeds. The difference between two such measurements will provide an indication of change.

The difference between two successive measurements indicates the general direction of change, but not usually, the meaningful amount of change. Interpretation of change or progress is achieved whenever the amount (and direction) of change can be related to (1) standards, (2) norms, or (3) meaningful units of measurement.

Standards

As used here, standards refer to those points along a continuum or to those discrete categories in a classification which permit the assigning of pupils to groups according to accepted definitions implying specified amounts of a characteristic. An obvious and frequently quoted example of a standard is the point along the continuum of temperature at which water freezes. Regardless of whether this point is expressed as Farenheit or Centigrade, its identification permits the classification of water as liquid or solid. Examples of standards from the educational realm include: can or can not tie shoestrings; can or can not read newspaper articles; can or can not perform long division; can or can not spell all the words in a given list; and has perfect or less than perfect attendance during a particular period. Whenever a standard is employed to obtain interpretability, precise definitions must be provided. For it is the element of communicability which contributes to interpretability. Thus a standard may be thought of as self-defining.

It should be noted that the adoption of a standard for gaining interpretability does not depend upon the amount of the characteristic displayed by a group at the start of an experimental experience. The group may or may not start at "zero amount of the characteristic." For example, at the start of a project, all children may be unable to tie their shoestrings or unable to read or unable to perform long division. In these instances, the group moves away from zero amount of the characteristic toward the specified standard. In other instances, the group or an individual may move from some baseline amount toward the standard. For example, at the start of the experimental experience, retention may be 90 per cent, while the standard adopted for the experiment may be 93 per cent for the following year. In all instances, the standard implies the desired outcome.

Norms

Norms, as used here, refer to numerical values that describe performance of specified groups, such as the standardization group for a test. Published norms for tests are often assumed to be representative of the nation as a whole. However, regional, local, or district norms may be desirable for comparison with specific groups. Whereas norms may be expressed in raw score form, more frequently age scores, grade placements, standard scores, stanines, or percentile ranks are used.

The reason for the use of norms in the measurement of educational outcomes is that most educational measures can not be interpreted until a comparison with some group is made. For example, it is not meaningful to say "Johnie spelled only ten words correctly on this list of twenty words." If Johnie is of average age in the third grade and the list of words was based on a high school spelling vocabulary, his performance might have been truly outstanding. On the other hand, if Johnie is of average age in the sixth grade and the words were from a third grade spelling vocabulary, his performance might be interpreted as very poor. By administering tests to large numbers of pupils comparable to those whose performance is to be interpreted, norms can be established so as to make any given performance meaningful through the process of relative comparison.

Units

Units are generally defined as increments of change expressed as meaningful amounts. Units may be expressed either by reporting norm data or simply stated as progress toward a standard. A few examples follow: number of times a pupil is chosen as a leader through sociometry, parental attendance at workshops, number of students seeking service of counselors, voluntary registrations for a second (advanced or continuing) course in study skills, and ratio of library cards issued to cards used.

Especially useful are percentage units. Although they do have limitations, as will be pointed out in a subsequent section of this document, they are easily understood by most people. For example, change in the percentage of a local group scoring below the national median before and after participation in a Title I project provides a meaningful combination of norms and units as follows:

> In school X it was found that 72 per cent of the local disadvantaged pupils at the start of the seventh grade fell below the national median in arithmetic at the 7.0 grade level. During the seventh grade, the local pupils participated in a Title I arithmetic self help project. When the pupils were tested at the end of the project year (nine months later) the percentage of the group scoring below the national median for the 7.9 grade level was 61 per cent.

Thus, a meaningful indication of the effectiveness of the project was obtained.

In general, the more carefully standards, norms, or units are chosen, the more meaningful will be the conclusion from an evaluation effort. Care should be taken to make sure that as many as possible of those who will use the outcome of the project can understand the standards, norms or units in which the results of the Title I project are expressed.

Testing the Culturally Deprived

Recently, the attention of educators has been focused, especially through Title I of the Elementary and Secondary Education Act, upon culturally disadvantaged children and the problems presented in evaluation of their educational progress. Nationally standardized tests are among the most widely used and most useful tools of educational personnel. How usefull tests may be with typical children is dependent upon the special training and sensitivity of personnel using them. When these tests are used with sub-cultural groups, the importance of thorough knowledge of the instrument and of the evaluation process is extended greatly. Unfortunately, there is no single all-encompassing and readily available reference to which test users can look for advice about testing children from minority groups. Recently a committee from the Society for the Psychological Study of Social Issues prepared a booklet outlining Guidelines for Testing Minority Group Children (See Deutsch, M. et al. in reference on page 95) in which are discussed the principal difficulties in using standardized tests with disadvantaged children. The three most important considerations include the hypotheses that (1) the tests may not provide reliable differentiation in the range of scores obtained from the disadvantaged group (2) the predictive validity of the tests for sub-groups may be quite different from that of the more frequently tested children, and (3) the validity of test score interpretation is very dependent upon a thorough understanding of the children with whom the tests are used. These three points are expanded and examples are provided in the booklet.

When scores of a special group of children fall within a narrow band of interpretative scores, particularly at the lower end of the scale, special norms for that group are often useful. If a large number of children have scores falling in the lowest 10 per cent of the national norms, local norms might be helpful in differentiating among the children. Procedures for computing percentile ranks are provided on page 58. Construction of local norms is advisable, however, only when it is obvious that student scores are more a function of their ability than of a chance factor. When the raw scores extend through a range considerably larger than the norm score range, reliability sufficient for local norm construction can usually be assumed.

In many instances the standing of the child in comparison with others of similar background is more appropriate than his standing among all children on whom the national norms are based.

When interest and personality inventories are used with groups whose culture is different from the standardization sample, even more caution should be exercised in interpretation than with measures of achievement and intelligence. With all tests that are assumed to penalize culturally different children, evaluators should see significant changes in scores as the culturally handicapping conditions are removed or alleviated. In many Title I projects the goal is to help the child become more culturally ready for the educational process which will in turn help overcome cultural disadvantages.

Lennon[1], in a recent speech, offered the following analogy,

> If we take a youngster who has suffered malnutrition over a period of years, who has not had the benefit of adequate health care, and put him on a scale, we may well discover that he is ten, or fifteen, or twenty pounds underweight. We do not then say the scale is biased because of the deprivation the child has suffered. We take this information as currently and accurately descriptive of an important fact about this child-a fact that can be used to his advantage in planning a program calculated to make up for the deficiencies in his earlier care. And if we do provide him with proper food and care, hopefully the scale will another time give us reassuring evidence of the success of our efforts. I suggest to you that this is the way of looking at a test score. The test is giving us a piece of information about a child's performance here and now, which information, if properly used, can be extremely helpful in planning the educational endeavors of the child.

---

[1] Lennon, R.T. "Testing and the Culturally Disadvantaged Child," available without charge from Harcourt, Brace and World, Inc. 1964.

As long as accomplishing the objectives measured by a standardized test is a worthy goal for children about whom we are concerned, we should use measuring instruments which reflect these achievements and find out where every child stands. Without the best possible evaluation instruments, we handicap ourselves in improving the learning activities.

Evaluation Designs

Evaluation involves assessing the extent and direction of change resulting from an educational experience. Evaluation designs are simply procedures which allow the experimenter to derive meaning from the amount and direction of changes which have occurred in a project group. Before meaning can be derived, however, consideration must be given to the units employed to record change, as well as to the selection of comparative data with which the observed change can be contrasted. Units are important in that they reflect the distance by which the project group falls short of some standard or norm. The standard itself reflects the value associated with the change that occurred. In general, the change in a project group will fall short of, be equal to, or exceed the specified standard. The function of evaluation designs is to facilitate the comparison and interpretation of the meaning of change resulting from an educational experience.

The foregoing comparison process implies that the evaluator has a variety of standards or norms from which to specify one for comparison with his project data. Actually, this is not always the case. Sometimes several sets of comparative data are readily available, while, at other times, even a single standard may be difficult to identify. Accessibility of standards or norms depends on many factors, including whether standardized or teacher-made tests were used, whether or not comparable groups not participating in the program were available, and whether or not repeated measures of the behavior concerned were made within the project prior to the start of the project.

Many different designs have been developed for purposes of educational evaluation, but most projects can be evaluated satisfactorily using one of the following types of design. Although these designs vary considerably in complexity, careful study of them reveals that they can be differentiated according to the source of the comparative data with which changes in the project group are contrasted. Two sources for comparison are immediately apparent: (1) data derived from within the project group itself and (2) data obtained from pupils or groups outside the project. Within each of these two major categories are three designs arranged in approximate order of complexity.

Comparison Data Derived Within the Project Group

When characteristics of a project group are measured at the start and at the end of a project, the initial measurement can be used as a point of comparison whenever more interpretative data are unavailable. In effect, change in these designs is a matter of moving away from the original position. The value of the change in such an instance must be derived from description only, from statistical analysis, or from an absolute standard, such as complete mastery of some task.

Design A. Title I Project Group Characteristics Compared with an Absolute Standard (100%). Basic data consist of simple numerical counts with the project data.

Example 1. Proportion of eligible tenth, eleventh, and twelfth grade students enrolled in a work study program.

Example 2. Proportion of former students enrolling in selected post-secondary educational programs.

Example 3. Proportion of parents accepting and participating in conferences with teachers.

Example 4. Proportion of students retained in school between the eleventh and twelfth grades.

In all these examples it is assumed that the standard is 100 per cent and that the closer the results are to 100 per cent, the more effective the experience has been.

Design B. Final Measurements of a Title I Project Group Compared with Initial Measurements. Basic data consist of raw scores, derived scores, ratings, or ratios within the project group.

Example 1. To evaluate a first grade reading project, a comparison can be made of scores earned on a standardized reading test administered in the fall, when the project began, and again at the end of the project year.

Example 2. To evaluate a program in which the pupil-teacher ratio has been reduced at each of the several grade levels and language arts and arithmetic supervisors have been employed, a comparison can be made of scores by the project group on each subtest in a comprehensive achievement battery of tests administered at the beginning and end of the project period.

Example 3. To assess change in social competence associated with a Title I project, ratings of observers made at the start and at the end of the project can be compared. Each pupil's post-test deviation from his pre-test position is then noted and overall differences are tested for statistical significance.

__Design C.__ __Final Measurement in a Title I Project Group Compared with Projected or Hypothesized Measurement Based on Past Progress of the Group.__ Basic data require at least one measurement made prior to the start of the project and one measurement at a later date.

Example 1. If an educationally handicapped group at grade level 4.0 is achieving at 2.0 at the start of a project, then the projection for achievement one year later will be 2.5 if proportional growth is assumed. Actual growth in the Title I group would then be compared with the "projected growth."

Example 2. Average intelligence quotients of 97, 93, and 89 were recorded in grades one, three, and five respectively for members of a Title I group. In the seventh grade, the projected average would be 85. (Often, the attainments of socially disadvantaged children have been recorded progressively further below average as yearly evaluations have been made.)

Example 3. Numerical values based on the physical condition of elementary school pupils participating in a Title I project had been 28, 29, and 28 on three consecutive years prior to the commencement of the project. At the start of the project, the mean score was 29, and at the end of the project, the mean score was 34. On the basis of the previous trends, the projected score would be 28. When the projected score is used as a point of comparison, the effectiveness of the project activity can be estimated by comparing obtained and expected outcomes.

In Design C, the results can be analyzed statistically by considering the obtained results as a sample and the projected standard as a population parameter.

Comparison Data Derived Outside the Project Group

In addition to the three foregoing designs in which data for comparison were obtained from within the project group, many designs exist for making comparisons with data external to the project itself. Three of these are described in this section.

Design D. Change in Title I Project Group Compared with a Designated Norm. Basic data are expressed as project group scores obtained at the start and at the end of a project and scores of a comparable group.

Example 1. When a nationally standardized achievement test is administered to a project group, the change in achievement of students in the project can be compared with expected change based on published norms. The percentage of the project group falling at or below the same point in the standardization group distribution (such as the median or 25th percentile) permits an especially meaningful comparison. Local conditions should be considered in specifying the most useful point of comparison.

Example 2. A norm is obtained from a different evaluation device administered to the same group. When ability and achievement tests are given to the same students, a comparison of pairs of standard scores or ranked positions can be made. Or, correlations between ability and achievement can be computed both at the start and at the end of the project period. In this instance, it is assumed that the norm for achievement is indicated by the ability displayed by the student.

Example 3. A norm is obtained from a different evaluation device administered to a different group. When achievement and ability tests are administered to a group, it is possible to make meaningful comparisons by converting the performance on the two tests to similar units, such as standard scores.

Design E. Change in Title I Project Group Compared with Change in Previous Class. Basic data are expressed in any type of unit. A previous year's class may be designated as the source for a standard, provided the pupils have comparable backgrounds and that the same evaluation devices were used to measure change at the appropriate times. The data from the Title I project are, then, compared with the data collected earlier.

Example 1. Following a nine-month reading improvement project, pupils in a project group showed a gain of 1.2 grade levels in reading. Gain for the previous year's class in the same school on the same test over the same period was .8 grade levels. Because of differences in the size of units along the grade placement scale, however, increments of gain expressed as months or grade levels should be interpreted very carefully. For example, two gains in reading score are comparable only if the pupils started from the same level. In general, standard scores provide a more meaningful comparison than grade placement units, since standard score units are more nearly equal throughout the range of the distribution.

Example 2. Scores on a vocabulary test increased from 65 to 84 for a Title I project group. This compares with a gain from 65 to 81 for last year's group. If desired, statistical significance can be determined by contrasted the two distributions.

Design F. Change in Title I Project Group Compared with Change in a Current Control Group. Basic data are expressed in any type of unit and consist of scores for two groups. A control group, as used here, is one similar to the Title I group with respect to the variables important to the specific activity or project, such as ability, socioeconomic level, etc. Ideally, the students are assigned randomly to the Title I and the control groups. However, such assignment is not necessary when it can be assumed that the students in both groups are equally prepared for the project's educational experience. The control group can be drawn from students outside the Title I project area who have the same type of deprivation.

Example 1. Both groups are required to take a comprehensive achievement test in October to establish a baseline and are required to repeat the test in May. Percentage in each group falling below the national median provides the comparison.

Example 2. Change in attendance record and holding power in a project school can be compared and contrasted with change in a control school during the same period.

Example 3. Change in kinds and severity of adjustment problems reported in a project group may be compared and contrasted with control group data obtained during the same period.

Need for Continuous Evaluation Regardless of Comparative Data

In addition to evaluating the final outcomes of an activity or project, continuous evaluation throughout a Title I project period is essential. Continuous evaluation means the process of making day-to-day observations and adjustments in the operation of a project to keep it functioning smoothly. Such observations and adjustments are usually a necessary part of the project, since it is literally impossible to anticipate the myriad of detailed decisions involved in a project until it is in actual operation. Obviously, these decisions and the solutions to unanticipated problems must occur within the framework of the overall objectives of the project. These decisions, often made on the basis of little or no available evidence, usually represent the best judgment of the teacher or project director in the light of the primary objective of the project. Reports of such decisions and problem solutions frequently have implications for utilizing the practice concerned in other situations.

Content of Evaluation

A comprehensive evaluation program will go well beyond testing for the mere acquisition of specific skills, facts, and knowledge of the cognitive domain. Comprehensive evaluation will extend into the measurement of the student's ability to interpret, to evaluate, and/or to extrapolate information to solve real problems. In fact, the purpose of American education goes far beyond student achievement in the cognitive domain to include concern for areas such as:

> the affective domain - attitudes, motivations, interests, adjustment, anxieties,
>
> social development - acceptance, recognition, belonging, leadership, interaction,
>
> physical development - general health and ability, speech, motor skills, dexterity, and
>
> academically related problems - reaction of employers, continuing professional development of teachers.

Attitude scales, personal evaluation, sociometric devices, speech pathology surveys, audiological surveys, physical examinations, participation in recreation program surveys, and many other instruments and devices may be used with professional observations of behavior to collect evidence about the total educational endeavor.

Desirable Characteristics of Tests

Although the evaluation of progress toward some objectives may require specially constructed instruments, progress toward most objectives can be evaluated with tests and scales already available. When selecting a measuring device for any evaluation purpose, it is helpful to recognize such criteria as those implied in the following statements for judging the satisfactoriness of a particular instrument.

A test should be _adapted to the range of the characteristic being measured_. Whenever a test is too easy or too difficult for the group involved, individual differences will not be apparent, especially at the extremes of the distribution. If a test is not adapted to the range of the characteristic being measured, there usually will be large proportions of the group tested who receive very low or very high scores.

A test should be _sensitive_. Closely related to the difficulty of a test is its sensitivity. This means that a satisfactory device should reflect relatively small individual differences among pupils throughout the range of the characteristics. A test which lacks sensitivity results in grouping pupils into coarse groups with insufficient differentiation among them. For example, if a ten-week project designed to improve the communication skills of sixth grade pupils is being evaluated, a highly sensitive device calibrated in very small units will be necessary to reflect the impact of the experience. Ten weeks is a relatively short period of time to modify already-existing skills. In general, the more specific the content of a test, and the larger the sample of behavior it represents, the more sensitive it will be.

A test should be _feasible_. Feasibility of tests involves such factors as cost, administration time, complexity of scoring, and availability of duplicate forms. For example, even though the administration of an individual intelligence test to pupils in a project might be desirable under circumstances, it may be necessary to use a group test instead because of time involved or the unavailability of trained examiners.

A test should be interpretable. Since most initial measures of educational and psychological characteristics of pupils are relative rather than absolute, some basis for making a score meaningful is essential. Interpretability refers to the extent to which meaningful comparison between or among groups being tested can be made. Such comparisons are usually facilitated by norms or standards obtained from previous administrations of the test. In general, most published tests have greater interpretability than locally constructed tests because of the variety of situations in which they have been administered.

A test should yield scores as free from error as possible. Just as all measures of physical characteristics contain errors, so do measures of educational and psychological characteristics. These errors of measurement are of two types,biased and compensating. Biased errors are those types of errors which do not cancel the effect of each error when measures are taken an infinite number of times by an infinite number of competent judges or when the instrument consists of an infinite number of items. Compensating errors, on the other hand, do have a tendency to cancel the effect of each error under the same conditions. Some compensating errors increase the score while others decrease the score.

Biased errors result from:

1. Failure to choose items which measure the desired function.
2. Failure to break down the characteristic to be evaluated to the place where it is homogeneous.
3. Failure to choose a good cross-section of items.

4. Failure to weight the items in the ratio of their importance.

5. Scorer imcompetency or dishonesty.

Compensating errors result from:

1. Failure to include an infinite number of items.

2. Failure to sample the reaction of examinees an infinite number of times.

3. Failure to use an infinite number of scorers in evaluating the response.

The extent to which a test is free from compensating error is an indication of its reliability or consistency. Thus, reliability can be thought of as the extent to which a test measures consistently whatever it measures. Although there are several procedures available for assessing reliability, the most frequently utilized are the coefficient of stability (same test administered to one group on two occasions); the coefficient of equivalence (two forms of a test administered to one group on the same occasion); the coefficient of stability and equivalence (two forms administered to one group on two occasions); and the coefficient of internal consistency (an estimate of reliability based on the interrelationship among subdivisions of the test).

Two of the more important aspects of reliability of measurement involve test length and difficulty of the items. Generally, the more reliable evaluations result from conditions in which a number of items measure each concept and the items are neither so easy that most are answered correctly nor so difficult that most are failed. In selecting tests, evaluators must assure themselves that the tests contain enough items which deal with concepts that were taught and that these items are comprehended by the students. When standardized tests contain too few items for a particular concept and/or the problems are too difficult, locally constructed tests should be used to supplement the evaluation.

The extent to which a test is free from both compensating and biased errors is an indication of its validity. Validity refers to the degree to which a test measures what it purports to measure. Four types of validity are widely recognized in educational and psychological measurement. Content validity refers to whether a test covers a representative sample of the behavior domain to be measured. Predictive validity refers to the extent to which a test will predict some future outcome. Estimates of predictive validity obviously must be obtained over a period of time. Thus, the relation between scores on a test and performance on a job is an index of predictive validity if the test is given at the time of selection and performance is reviewed at a later date. Concurrent validity refers to the relationship between scores on a test and some indication of status independent of the test itself such as scores on another test, ratings by teachers, etc. Concurrent validity can not be predictive since both measures are obtained at about the same time. Construct validity refers to the extent to which a test measures a theoretical construct or definition of some characteristic. Construct validity has its basis in the theoretical foundation of the trait being measured. The results of factor analyses, internal consistency of the test itself, and the differentiation among groups of children at successive age levels are data used to demonstrate construct validity of a test.

Regardless of the validity and reliability data provided with a published test, there is no substitute for a careful and thorough review of the content of a test in relation to the objectives of the project involved. In addition, pretests and pilot studies will be of considerable assistance in choosing appropriate tests for the evaluation of educational outcomes.

The Use of Standardized Tests

Whenever possible, "objective measures of educational achievement" will be used for the evaluation of Title I projects. In most instances this will mean nationally standardized tests. Extreme care must be taken, however, to assure that the standardized tests are valid measures of the objectives. For example, if an objective involves spelling words correctly while writing, the typical standardized spelling test wherein the pupil is asked to judge the correctness of printed words will not suffice for evaluation of this objective. This is another way of saying that the evaluation devices must be direct outgrowths of the objectives.

Supplementary Evaluative Techniques

Some local educational agencies may have difficulty reporting significant changes in educational attainment for a project group, because the nature of the project is such that conclusive results will not be available for two or three years. In the interim, however, individual cases may serve to demonstrate meaningful increases in educational attainment.

Case Studies: Appraisals by teachers or Title I project directors of changes in attitudes and behavior must be well documented to be reliable. Each teacher in the course of observing and testing his students, as well as in numerous other ways, acquires many important facts about them. For reporting purposes under Title I, it would be helpful if the accumulated facts were presented in terms of some specific aspect of the participant's development.

Such appraisals should be based on more than mere "feeling," for a "feeling" can not be replicated or checked. This is not to say that a "feeling" is not useful, but that, to be of convincing value, it must be supported by a carefully marshalled, detailed description of cases and observations. Observations are a more reliable evaluative device if made by skilled "outside" observers not connected with the project or program.

Anecdotal Records: Anecdotal records may be employed by teachers and counselors to evaluate Title I projects. An anecdotal record consists of an accumulation of a series of observations on a significant aspect of a student - his leadership qualities, reading achievement, socialization. The individual report of each incident should be a brief, clear, objective statement of what took place. Interpretation or recommendations may be included, but on separate sections of the anecdotal card or form. The observations must be objectively recorded and taken at periodic intervals in order to show individual development. Teachers and other project personnel may need to train themselves to observe incidents and to record them at a later time.

Related Devices: Attitudinal scales, personal evaluations, teacher rating forms, pupil self-rating inventories, audiological surveys, physical examinations, participation in recreation program surveys, and many other instruments and devices may be used along with professional staff observations to collect evidence about the total impact of Title I projects.

Several examples illustrating the content of items for evaluating attitudes and reactions are shown in the Appendix. Reference to a textbook is necessary before constructing such forms locally. Attitude scales take many forms and can be constructed to assist in the evaluation of many characteristics. The first page of scales to measure attitude toward methods used in geometry class, toward an English teacher, and toward general teacher characteristics, as used by the authors in a study of the Reactions of High School Students to Television Teachers,[1] can be found on pages 111-114. It is difficult to obtain high reliability with a few items. Consequently, each object, such as method of teaching or teacher characteristics, should have a number of items employed in its measurement. Following these scales is a TV checklist used as an example in the Connecticut guidelines.[2] Such checklists take many forms and employ not only words and phrases, but also sentences and paragraphs which are checked to denote the presence or absence of certain conditions.

A useful tool for recording observations is the rating scale. Ratings can be quantified more easily than the anecdotal record because the rater assigns a value along a continuum for each trait. Usually a five point scale is used, such as in the Teacher Characteristic and Activities Participation Scale found in the Appendix and in the selected items from a scale which follows:

---

[1] Neidt, C.O. and French, J.L. Reactions of High School Students to Television Teachers, Lincoln, Nebr.: The University of Nebraska, 1958.

[2] Connecticut State Department of Education. Evaluating Programs Approved Under Title I Public Law 89-10. Hartford: The Department, 1965.

optimal	good	average	detrimental	seriously detrimental

Attitude Toward the Task

serious	playful
enthusiastic	indifferent
works with eagerness	performs reluctantly
guesses without fear	refuses to answer

Attitude Toward His Own Performance

recognizes errors	does not recognize errors or failures
redoubles effort when puzzled	gives up easily
persists when failure apparent	attempts to change to easier task

The items above are only seven of 37 used to measure the objectives in a specific situation. Rating scales should be tailored to each situation rather than borrowed from someone else. An example of a scale which provides more assistance to the rater in determining the point along the continuum appropriate for a specific performance follows:

Attention

Fixes attention on each card	Occasionally needs to be told to look at cards	Must be told to look at cards about every third time	Must be told to look at most cards	Attention must be directed to each card

An Illustrative taxonomy

From the foregoing comments it is apparent that many different techniques are available for evaluating projects under Title I. Choice of a particular technique will depend upon many factors, including the objective of the project, the availability of published instruments, and the sophistication of persons using the technique during the project period. Although several classifications of techniques have been prepared for publication in textbooks and in general references by authorities in the field of evaluation, useful classifications of this type are those prepared by the Connecticut State Department of Education for publication in their guidelines.[1] The authors of the Connecticut guidelines point out that this is not a complete list of either outcomes or techniques; rather, it is a concise and illustrative listing. The Connecticut classification is as follows:

1. Subject matter and skill achievement

   appropriate standardized tests
   teacher-made objective tests
   teacher-made performance tests

2. Changes in attitude

   observation (particulary by outside observers)
   questionnaires, to be answered by pupils or parents
   rating scales
   dropout counts (changes, comparisons)
   records of parent involvement in school-sponsored projects
   case studies
   anecdotal records
   attendance records
   records of participation in an activity

---

[1] Connecticut State Department of Education. *Evaluating programs approved under Title I of Public Law 89-10*. Hartford: The Department, 1965 p. 17-18.

3. Interest

    questionnaires
    attendance records
    case studies
    anecdotal records
    dropout counts
    records of parent involvement
    tabulations (such as average number of books read per pupil)
    rating scales
    check lists

4. Ideals

    anecdotal records
    observation
    pupils' writings

5. Ways of thinking

    appropriate standardized tests (rare)
    teacher-made tests
    rating scales
    pupils' writings

6. Work habits

    observation
    anecdotal records
    rating scales
    check lists

7. Personal and social adaptability

    dropout information
    attendance records
    anecdotal records
    rating scales
    pupils' writings
    sociograms
    case studies

Since attempts should be made to evaluate each objective, imaginative, innovative, and creative thought are needed in the planning stage of the evaluation. When thoughts are not directed to evaluation prior to initiating the project, it is often impossible to collect "objective measurements" of crucial elements indicating important change. The importance of early planning of evaluation procedures can not be overemphasized. Consultation with educational researchers located in other school systems, institutions of higher education, regional educational laboratories, State departments of public instruction, or elsewhere will be effective at this stage and will help to determine when continued consultation is necessary. In most school systems, funds will have to be designated for personnel training and the employment of consultants.

Sources of Project Data

Although most of the Title I project data will come from pupils, information will sometimes be sought from parents, school personnel, and others in the community. Opinions and attitudes of parents will often be particularly important. While questionnaires and other survey instruments may be used, their effectiveness is limited because they elicit verbal reports rather than actual behavior. Attendance at school activities, participation in conferences, and other examples of adult behavior, such as watching certain television programs, providing a quiet place for homework, and using library facilities, will provide indices of opinion through actual choices of behavior.

Professional judgments of teachers, specialists, and supervisors about children or about a project can be obtained with specially constructed rating scales. One staff attempted to assess faculty attitudes toward various phases of their in-service program by using the "paired-comparison" technique. Abbreviated instructions for a portion of their instrument appears in Figure 1. A rating scale as previously described could have been used but the technique illustrated in Figure 1 reduces the halo effect found in some raters and forces a ranking of the various activities. Pooling the results of raters will suggest the relative value of the various activities.

Effectiveness of in-service teacher development projects can be evaluated by observing the change in a pupil's performance on achievement tests and by procedures dictated by objectives, such as conducting, at regular intervals from the beginning to the end of the project, an analysis of the

236-150 O-66—4

Evaluation of Informal Seminar

The following is a list of activities in which we have engaged this year. The activities are grouped in pairs. You are to read and think about each pair and decide upon the one that has been most helpful to you. You must make a choice between each pair. Place an X on the line beside the statement of your choice. It is possible that you liked both activities very well. Even so you must decide which one contributed more to you and your thinking about in-service projects. Also, you will find some pairs that were of little value to you. You must decide between each pair and put an X on the line beside the one which contributed more to you.

Discussion of Literature	___	___	Discussion of Visitations
Planning Sessions	___	___	Recapitulation Seminars
Formulation of Objectives	___	___	Formulation of Philosophy
Discussion of Literature	___	___	Planning Sessions
Formulating of Philosophy	___	___	Discussion of Literature
Discussion of Visitations	___	___	Recapitulation Seminars
Formulation of Philosophy	___	___	Recapitulation Seminars
Recapitulation Seminars	___	___	Discussion of Literature
Formulation of Objectives	___	___	Planning Session
Discussion of Visitations	___	___	Planning Session
Formulation of Objectives	___	___	Discussion of Visitations
Discussion of Literature	___	___	Formulation of Objectives
Planning Sessions	___	___	Formulation of Philosophy
Formulation of Philosophy	___	___	Discussion of Visitations
Recapitulation Seminars	___	___	Formulation of Objectives

Figure 1. An example of the "paired comparison" technique for obtaining faculty ratings of an in-service education program.

interaction of pupils and teachers.

The community offers many sources of data that can be used to supplement the evaluation of objectives in the cognitive, affective, and other areas. The library is a good source for providing several kinds of data. Evidence pertaining to literature, social studies, vocational fields, and other fields can be collected from the number and types of books borrowed by students and/or recent students. Police records can be studied to supplement the evaluation of some projects by examining police contacts with project and non-project students from the same district. Participation in community and other non-school activities can be reviewed in an effort to learn more about the use of leisure time. Attendance at post-secondary school educational institutions can be related to attitude and interest, as well as to achievement studies.

Throughout this document, reference has been made repeatedly to assessing the extent to which objectives of a Title I project have been achieved. This is the central core of evaluation and should be the primary consideration. It is important to recognize, however, that a project may have an influence on behaviors other than those specified in the project objectives. For example, if a project involves extensive emphasis on reading skills, the question can also be raised (in addition to those relating to improved reading skills) "what happened to the arithmetic skills while the reading program was going on?" Complete evaluation involves going outside the project itself to seek descriptions and judgments from many sources so that the total impact of a new educational experience can be assessed.

Division of Groups

When objectives in the cognitive domain are being evaluated, it may be desirable to consider a subclassification of pupils according to sex, ability level, and/or achievement level. Such a practice is suggested by several previous research studies which reveal that certain educational practices result in substantial gains for some groups but not for others (i.e., girls but not boys or students with IQs of 110 and above but not students with IQs of 90 and below). Such subclassification of experimental pupils often provides insight into educational practices which might otherwise be "masked" through pooling divergent groups together in an analysis of evaluative data. In a recent study of high school dropouts with IQs of 110 and above, it was observed from self-concept data that 38 per cent of the dropouts believed they were "hard headed." A further look at the data revealed that 48 per cent of the male dropouts believed they were "hard headed" and 28 per cent of the females checked that item. In data from a group of students with equal IQs from the same neighborhoods who stayed in school through graduation, it was observed that 43 per cent of the boys and 25 per cent of the girls believed they were "hard headed." The fact that 38 per cent of these dropouts believed they were "hard headed" appeared to be significant until looking at responses of boys and girls separately and until looking at control group data. In the neighborhoods studied, about half of the boys and a quarter of the girls, whether they stayed in school or dropped out, believed they were "hard headed." Group statistics often mask important information about sub-groups. In many aspects of the dropout study mentioned above, the data were quite different for married and unmarried female dropouts. To look at variables in the affective domain for female dropouts without further subdivision is a mistake.

As is true of evaluation designs and the selection or development of evaluative instruments, local innovation and imagination must be employed in identifying appropriate sources of data with which to evaluate progress in attaining objectives.

Preparation of the Proposal

Detailed procedures for evaluation should be described in each proposal submitted under Title I. The evaluation procedure should be selected during the planning stage to insure the proper collection of necessary data both prior to and during the project. Preparation for the evaluation of each project should begin by listing the need or needs, the objectives arising therefrom in behavioral terms, and parallel listings of instruments to be used in evaluating progress toward each objective.

Although defining activities and educational experiences that will fulfill the stated needs will not enter directly into the evaluation activities, it is obvious that such activities must grow from the objectives. Further, if the activities are well chosen and carried out, the outcomes will be achieved. If the activities are not well chosen and carried out, evaluation will not be necessary and the time, energy and money spent on the project will be wasted. This is another way of saying that educational activities must bear a rational relationship to the need the project is designed to fill.

A helpful chart for setting up an evaluation plan has been reproduced in Figure 2. The examples used in Figure 2 are fictitious and incomplete and are used here for illustrative purposes. Examples involving objective test data were not included because they are more readily understood. The plan for evaluation should include a brief description of the instruments or techniques which will be used to measure each outcome, the schedule for applying the techniques, and any other information that might be of use in describing the nature of the techniques and procedures.

Observed Need	Illustrative Objective	Behavioral Outcome	Evaluation Techniques	Source	Schedule	Sample	Remarks
Reading ability below grade level	To improve work attack skills	Pronounces words not previously read	Teacher-made tests	Teacher	Once a month	All children Gr.1-6	Compare results with standard achievement test data
Too many students not graduating from high school	To retain able students in school until graduation	Higher proportion of students graduate each year	School records	Principal's office	End of each school year	All youth	Do achievement test data suggest retention was helpful? Review records of youth in work study program.
Students reflecting poor attitude toward school and activities	To improve student perception of value of of classes and activities	1.Students do not deface property 2.Students attend actvities regularly 3.Students say they like school	1.Survey of equipment, materials, and building 2.Count of those in attendance at activities	1.Janitorial staff 2.Counselors and chaperones 3.University research group	End of each school year	1.All buildings etc. 2.Sample designated in fall 3. All youth	1. Compare with previous records 2. Compare each of two groups and types of activities 3. Compare data by curriculum and by sex. Check reliability with research group.
Twelve children with poor perception	To improve visual perceptual organization	Children notice fine distinctions among common things		1.Teacher 2.Psychologist	1. At least every other week 2.Beginning and end of year	Twelve designated children	1. & 2. Children designated are those most in need of help at this grade level. 2. Check reliability with another local sample.

Figure 2. Chart illustrating the relationship among the phases of the evaluation of selected portions of hypothetical Title I projects.

This last point could include limitations imposed by certain techniques, the relation of one technique to other outcomes, why a technique is used with a small sample, etc.

The form in which the proposal is finally written should comply with guidelines issued by each state. The chart suggested by Figure 2 will help with organizing pertinent information for the state form and assure that the evaluation will be appropriately included in the proposal.

Analysis of Data

Since four levels of responsibility are involved in evaluation under Title I (local, State, Federal and National Council), the analysis of data collected at each level will vary according to the purpose and the scope of the level concerned. Whereas evaluation data at the local level will be specific to individual activities and projects, the large amounts of data involved at State and Federal levels will require extensive summarization and synthesis. To assure that local evaluation data are assembled and analyzed in an efficient manner for reporting purposes, the United States Office of Education and the various State Offices of Education will provide guidelines for use by the local schools. The data and analyses requested by State and Federal agencies will, in general, represent only minimum requirements for assessing the impact of Title I. The local agencies, therefore, may wish to make considerably more extensive analyses of results than are required for reporting purposes. Such a practice is highly desirable since greater understanding and communication result from thorough analyses of the data collected.

Several procedures for analyzing evaluation data are described here in order to illustrate those processes useful in reaching meaningful conclusions and in making sound decisions about specific educational practices. In the discussion which follows, only a rudimentary knowledge of statistical methodology is assumed.

Raw Scores

Evaluation data in education represent observations of the behavior or other characteristics of some basic unit or case. The case may be a pupil, a parent, a teacher, a school system or any defined element. In any particular project under Title I, one or more groups of cases will be involved.

Frequently, a project will involve several groups such as classes or schools. But each group is composed of cases.

Observation of the characteristics of cases is facilitated by the use of measuring instruments which reflect amounts and kinds of individual differences with respect to designated characteristics. The assembled observations of cases being studied constitute a distribution. Thus, if differences in the spelling ability of 25 pupils in a third grade class are to be reflected, these differences can be shown with the use of a spelling test to describe each child's ability. In this example, the child is the case, the class is the group, spelling ability is the characteristic, and the spelling test is the measuring instrument. When all children have taken the test, assuming that it is a valid test appropriate for their ability level, they can be differentiated one from another according to the number of words spelled correctly. The number of units of some characteristic possessed by any case, expressed in terms of a specific measuring instrument, is a raw score. In this example, the raw score is the number of words spelled correctly.

Raw scores, however, are meaningful in and by themselves only when the zero point on the measuring instrument coincides with zero amount of the characteristic. Thus, height in inches is a meaningful score, since zero height corresponds to zero inches. Such is not the case with most measurements of educational and psychological characteristics. A single raw score on a spelling test is not meaningful since zero on the test does not necessarily mean a complete absence of spelling ability. Likewise, until the length of the test (number of words to be spelled), and its difficulty are known, no real meaning can be attached to a raw score. In relative rather than absolute measurements, the raw score becomes meaningful or interpretable only as it is transposed to some measure of relative position through comparing it with the scores obtained from applying the same measuring instrument to an appropriate group. The group to which the measuring instrument is administered so as to permit subsequent comparisons is referred to as a standardization group.

Derived Scores

Several methods of expressing the relative position of a raw score within a group of scores have been developed. These are referred to as derived scores. Derived scores are more meaningful than raw scores since they can be readily interpreted in terms of the group of cases on which they are based. The most frequently utilized derived scores are grade equivalents, age equivalents, percentiles, standard scores and Standard T scores. Each of these is described in the following paragraphs.

Grade equivalents. A widely-used procedure for gaining interpretability of achievement test raw scores is that of converting them to grade equivalents. A grade equivalent for a given raw score is the real or estimated average (mean or median) grade level of pupils in the standardization group who have obtained that score. For grade equivalent purposes, the school year can be defined as consisting of either nine or ten months and the interpretation is: a grade equivalent of 6.3 is that performance reflected by the average pupil during the third month of the sixth grade level.

As was indicated previously, the grade equivalent may be a real or estimated average. In the standardization process, if some raw scores are not obtained, grade equivalents may be estimated by interpolation or they may be extrapolated beyond the range of the actual raw scores. Most publishers describe in the test manual the computational procedure used in obtaining reported grade equivalents.

Age equivalents. Age equivalents correspond closely to grade equivalents in that an age equivalent is the real or estimated average age of pupils obtaining a given score. Age equivalents had their origin in the concept of mental age developed for expressing performance on intelligence tests. In the original mental age concept, however, the test constructor usually chose a sample such that groups of children at specified ages only, for example, children from 4 years and 11 months through 5 years and 1 month, would be classified as five years of age, etc., were examined. In the age equivalent concept, the test is administered to a large sample of children of varying ages and the median or mean age of those obtaining each raw score is computed or estimated.

Percentiles. Percentiles are points in the distribution of scores which divide the cases into one hundredths. Thus a percentile is a point in a distribution below which fall the per cent of cases indicated by the given percentile. The first percentile is that point in a distribution below which one per cent of the cases lie; the second percentile is that point below which two per cent of the cases lie; the second percentile is that point below which two per cent of the cases lie; and the ninety-ninth percentile is that point below which 99 per cent of the cases lie. The distance between two consecutive percentile points is a percentile rank. Any value appearing at the first percentile or below is assigned a percentile rank of one; between the first and second percentile a percentile rank of two; and so on, and any value appearing above the ninety-ninth percentile has a percentile rank of 100. As is true of percentages, the computation of percentiles is most meaningful when the number of cases on which they are based is relatively large.

In addition to specific percentiles and percentile ranks, other points and ranges in a distribution may be cited. The median corresponds to the 50th percentile and any score above this value is said to fall in the upper half of the distribution; or conversely, in the lower half. The two quartiles corresponding to the 75th and the 25th percentiles (along with the median) separate the distribution into quarters. The nine deciles corresponding to the 10th, 20th, 30th, etc., percentiles separate a distribution into tenths. Less frequently encountered are terciles and quintiles. It should be apparent, however, that all the points mentioned here are merely extensions of the percentile concept.

Computation. The method of computing percentiles and percentile ranks is illustrated as in Table 1 The data are first arranged in a frequency distribution (See "distribution" in Glossary) with the highest or best scores at the top of the distribution. The simplest frequency distribution is one having an interval of one. In such a distribution, all the scores are arranged from highest to lowest in an array and the number of pupils obtaining each score (frequency) is tabulated. The percentile for each interval is determined by dividing each cumulative frequency entry by the total number of cases in the distribution and multiplying that value by 100. Percentile ranks are obtained by rounding any decimal fraction of a percentile upward to the next whole number. By arranging the scores with the best scores at the top and the lowest at the bottom and by computing the cumulative frequencies from the bottom of the distribution to the top, a percentile rank of one represents a low score. In the example shown in Table 1 a raw score of 22 corresponds to a percentile rank of 93 and a raw score of 13 corresponds to percentile rank of 29.

Table 1

Computation of Percentiles and Percentile Ranks

Raw Score Interval	Frequency	Cumulative Frequency	Percentile	Percentile Rank
27	2	337		100
26	4	335	99.41	100
25	3	331	98.22	99
24	7	328	97.33	98
23	8	321	95.25	96
22	11	313	92.88	93
21	18	302	89.61	90
20	17	284	84.27	85
19	24	267	79.23	80
18	29	243	72.11	73
17	31	214	63.50	64
16	30	183	54.30	55
15	28	153	45.40	46
14	30	125	37.09	38
13	17	95	28.19	29
12	28	78	23.15	24
11	12	50	14.84	15
10	14	38	11.28	12
9	9	24	7.12	8
8	6	15	4.45	5
7	3	9	2.67	3
6	0	6	1.78	2
5	2	6	1.78	2
4	1	4	1.19	2
3	3	3	.89	1

Performance within a group can be readily described by citing the percentage of the group falling at or below any given score. If a test has been administered to two or more groups, rough comparison can be made by citing the percentages within each group obtaining a given score or below.

Although percentiles are easily interpreted, they do have the disadvantage of inequality throughout the distribution. Thus, the first percentile rank usually represents a wider range of the characteristic measured than the forty-fifth or the sixtieth. For this reason, percentile ranks should not be averaged and should be used as descriptive values only. The inequality of percentile ranks is discussed in greater detail in subsequent paragraphs.

Standard Scores

A standard score represents any one of a group of derived scores in which the position of any individual is expressed in terms of the number of standard deviations which his score lies away from the arithmetic mean of the group. An understanding of standard scores requires thorough knowledge of the computation of the mean and of the standard deviation.

The mean (arithmetic mean) of a distribution of scores is the sum of the scores divided by their number. Thus, in the following distribution of four scores, 6, 8, 8, 6, the mean is 28/4 or 7. If X represents any score in the distribution, N the number of scores and Σ "the sum of," the computation of the mean can be expressed as

$$\text{Mean} = \frac{\Sigma X}{N}$$

It is customary to use either of two symbols, M or $\overline{X}$, to represent the mean.

The standard deviation of a distribution is defined as the square root of the mean of the squares of the individual deviations from the mean of a distribution. In computing the standard deviation, the mean is first subtracted from each score in the distribution to obtain the deviation of each score from the mean. The deviations are squared, summed, divided by the number of scores, and the square root is extracted. The symbol for the standard deviation is S.D. or σ (lower case Greek sigma). The formula for the standard deviation is

$$\text{S.D.} = \sqrt{\frac{\Sigma(X - \overline{X})}{N}} \quad \text{or} \quad \sqrt{\frac{\Sigma x^2}{N}} \quad \text{where } x = (X - \overline{X})$$

where X = any score in the distribution
$\overline{X}$ = the mean of the distribution
and N = the number of scores in the distribution

The computation of the standard deviation is illustrated with the following nine scores.

X (Score)	x $X - \overline{X}$	x^2 $(X - \overline{X})^2$
9	+4	16
8	+3	9
7	+2	4
6	+1	1
5	0	0
4	-1	1
3	-2	4
2	-3	9
1	-4	16
45	0	60

Obviously, such a procedure requires the availability of the percentages of area under the normal curve corresponding to standard deviation distances away from the mean. Such values are published in most statistics texts and have been reproduced in abbreviated form in a later section of this document. The detailed computation of Standard T scores is discussed on page 65 under the heading "Converting Evaluation Data to Standard T Scores."

The relationship between the standard deviation and the area under the normal curve is shown in Figure 3. From this figure it can be seen that, if the total area encompassed by the curve is represented as 100%, ordinates erected at a distance corresponding to the mean plus and minus one standard deviation will encompass 68.26% of the area. Since 50% of the area falls above and below an ordinate erected at the mean, it follows that a point at +1.0 standard deviation above the mean corresponds to a percentile of 84.13 and that a point -2.0 standard deviation distances below the mean corresponds to a percentile of 2.18. Either standard scores or percentiles for a normal distribution can be inferred from the known areas. It is the concept of known areas under the normal curve in relation to standard deviation distances away from the mean upon which Standard T scores is based.

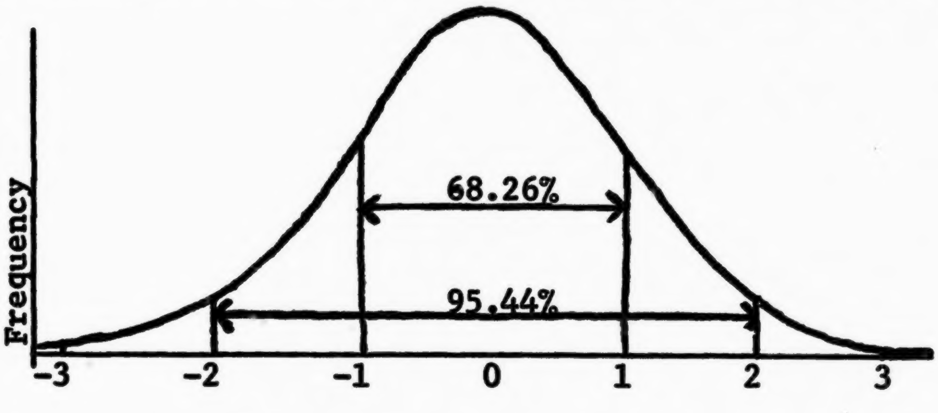


Figure 3. The Normal Distribution.

$$\overline{X} = \frac{\Sigma X}{N} = \frac{45}{9} = 5$$

$$\sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}} = \sqrt{\frac{60}{9}} = \sqrt{6.67} = 2.58$$

Although there are several modifications of standard scores in wide use, all are variations of the following definition:

$$\text{standard score} = \frac{X - \overline{X}}{\sigma} \text{ or } \frac{x}{\sigma} \text{ where } x = X - \overline{X}$$

where X = any score in the distribution
$\overline{X}$ = the mean of the distribution
and σ = the standard deviation of the distribution

In the foregoing distribution, the standard score corresponding to a raw score of 8 is 1.36 since

$$\frac{8-5}{2.58} = 1.36$$

It follows that if there are approximately as many scores above as below the mean of a distribution, approximately half of them will carry a negative sign if the foregoing formula for computing standard scores is used. In addition, decimal fractions will be involved frequently whenever the mean or the standard deviation are whole numbers and precise description is desired. To avoid both negative signs and decimal fractions, it is common practice to multiply the numerator by a factor such as 10 and to add a constant to the quotient. This forces the distribution of standard scores to have a mean and standard deviation equivalent to the constant and the factor respectively.

For example, the following procedures are widely used.

Mean	Standard deviation	Formula
50	10	$\frac{10(X - \overline{X})}{\sigma} + 50$
100	20	$\frac{20(X - \overline{X})}{\sigma} + 100$
100	15	$\frac{15(X - \overline{X})}{\sigma} + 100$
500	100	$\frac{100(X - \overline{X})}{\sigma} + 500$
12	3	$\frac{3(X - \overline{X})}{\sigma} + 12$

All the foregoing are variations of standard scores.

It should be noted that changing all raw scores in a distribution to standard scores of this type does not change the shape of the distribution. If the distribution is not symmetrical in raw score form, neither will it be symmetrical in standard score form.

Since the unit of measurement utilized throughout the range of raw scores is constant, distances between designated standard scores are comparable throughout the range. Thus, the distance between standard scores of .5 and 1.0 is the same as that between 1.0 and 1.5. It will be recalled that this was not true of percentiles.

Standard T Scores

Since relatively little evidence can be found to the contrary, it is usually assumed that most educational and psychological characteristics follow the "normal distribution" in the general population. The normal distribution is a symmetrical, bell-shaped curve, high in the center. From the center, the height decreases slowly at first and then more rapidly until it is about three-fourths of its original height, after which the rate of decrease becomes smaller and smaller until it is imperceptible. The mathematical equation for a curve corresponding to the normal distribution is known and has provided many useful applications of the normal distribution concept in educational and psychological measurement. Among these is the Standard T score. The Standard T score is a type of standard score with a mean of 50 and a standard deviation of 10. It differs from the customary standard score, however, in that its distribution is forced to follow the normal curve. This is accomplished by first transposing the raw scores in the original distribution into percentiles and then converting these into equivalent standard deviation distances away from the mean of a normal distribution.

Equality of Units

Most published tests are accompanied by tables for converting raw scores to one or more types of derived scores. Of the various types of derived scores usually available with published tests, the Standard T score based upon the normal curve with a mean of 50 and a standard deviation of 10 is probably the most useful in the majority of situations. As a method of expressing evaluation data Standard T scores have the advantage over easily understood percentiles since the units in standard T scores are more nearly equal throughout the range than are those of percentiles. Thus, the differences of 5 units each between the percentiles of 5 and 10 and 50 and 55 are not comparable because they represent different amounts of the characteristic being measured. On the other hand, the differences of 5 units each between the Standard T scores of 30 and 35 and 55 and 60 are comparable. Standard T scores can be averaged and subjected to statistical analysis, whereas percentiles should not be. If the percentile corresponding to the mean raw score of a group is desired, the mean raw score should be computed first and then converted to a percentile rather than converting each raw score to a percentile and averaging these.

Unequal units also are characteristic of age and grade equivalent norms, but usually to a lesser degree than percentiles. In age and grade equivalents, the assumption is often made that growth throughout the school year progresses at a constant rate. Such an assumption which permits the monthly units to be interpolated between successive grade or age levels when the test is standardized is seldom a valid one. Likewise, a unit of one month of change in the first grade is hardly equivalent to a unit of one month of change in the tenth grade. Tentative comparisons can be made however, using age and grade equivalent norms between groups at the same level within the same subject matter area.

Descriptive Units

In some evaluation situations measuring instruments employing descriptive units will be utilized. Descriptive units are qualitative terms used to describe amounts of some characteristic. Examples include superior, excellent, good, fair, poor or A B C D E. Whenever these descriptive units are arranged in order so as to form a scale, they can be converted to numerical scores by assigning code values to each descriptive term in the sequence. Although precise statistical techniques for converting descriptive units to numerical units have been developed, the assignment of consecutive integers to the units is usually quite satisfactory. Thus, course marks become A = 4, B = 3, C = 2, etc., or Superior = 5, Excellent = 4, Good = 3, etc. Once the descriptive units have been converted to numerical units, they can be treated as any other raw scores.

Converting Evaluation Data to Standard T Score

As indicated earlier, one of the most appropriate procedures for analyzing evaluation data is the use of Standard T scores. The recommended procedure for using Standard T scores is, first, to convert mean raw scores to a corresponding percentile and, then, to convert the percentile to a Standard T-score. Standard T scores can be reported for pre- and post-test administrations or for any other comparisons desired. A conversion table for transposing percentiles to Standard T scores is shown in Table 2.

Table 2

Table for Converting Percentiles to Standard T scores.

Percentile-Standard T	Percentile-Standard T	Percentile-Standard T	Percentile-Standard T
1 - 26.7	26 - 43.6	51 - 50.3	76 - 57.1
2 - 29.5	27 - 43.9	52 - 50.5	77 - 57.4
3 - 31.2	28 - 44.2	53 - 50.7	78 - 57.7
4 - 32.5	29 - 44.5	54 - 51.0	79 - 58.1
5 - 33.5	30 - 44.7	55 - 51.3	80 - 58.4
6 - 34.5	31 - 45.0	56 - 51.5	81 - 58.8
7 - 35.2	32 - 45.4	57 - 51.8	82 - 59.2
8 - 36.0	33 - 45.6	58 - 52.0	83 - 59.5
9 - 36.6	34 - 45.9	59 - 52.3	84 - 59.9
10 - 37.2	35 - 46.2	60 - 52.5	85 - 60.4
11 - 37.7	36 - 46.4	61 - 52.8	86 - 60.8
12 - 38.3	37 - 46.7	62 - 53.0	87 - 61.3
13 - 38.7	38 - 47.0	63 - 53.3	88 - 61.7
14 - 39.2	39 - 47.2	64 - 53.6	89 - 62.3
15 - 39.6	40 - 47.5	65 - 53.8	90 - 62.8
16 - 40.1	41 - 47.7	66 - 54.1	91 - 63.4
17 - 40.5	42 - 48.0	67 - 54.4	92 - 64.0
18 - 40.8	43 - 48.3	68 - 54.7	93 - 64.8
19 - 41.2	44 - 48.5	69 - 55.0	94 - 65.6
20 - 41.6	45 - 48.7	70 - 55.3	95 - 66.5
21 - 41.9	46 - 49.0	71 - 55.5	96 - 67.5
22 - 42.3	47 - 49.3	72 - 55.8	97 - 68.8
23 - 42.6	48 - 49.5	73 - 56.1	98 - 70.3
24 - 42.9	49 - 49.7	74 - 56.4	99 - 70.5
25 - 43.3	50 - 50.0	75 - 56.7	

To illustrate the use of Table 2 , the following hypothetical data will be used:

Case	Pretest Raw Score	Post Test Raw Score
John Doe	23	32
Richard Noe	20	28
Jane Moe	26	28
Ed Poe	23	36
Sum	92	124
Mean (9 2/4)	23	(12 4/4) 31
Percentile from manual	33	42
Standard T score (Table 1)	45.6	48.0
Change Between Pre and Post		48.0 - 45.6 = 2.4

Change occuring between test scores administered prior to and at the end of a Title I project then can be either shown in tables for various groups or plotted graphically. Examples of the use of graphs to reflect change in pupils in a Title I project are shown in Figure 4, taken from "Instructions for Title I Evaluation Reports" prepared by the Colorado State Department of Education.[1]

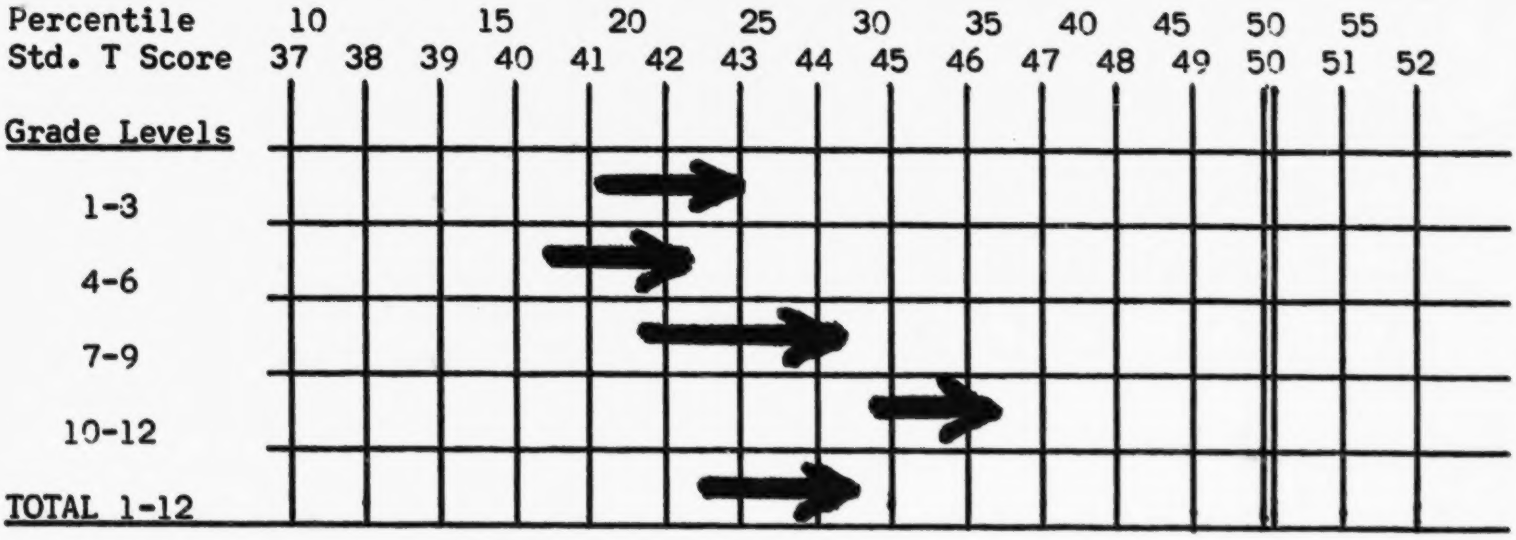
---

[1] Colorado State Department of Education, "Instructions for Title I Evaluation Reports" 1965-66, Denver, Colorado, April 1966.

RESULTS OF ACHIEVEMENT TESTING
TITLE I, ESEA - PROJECT NO. 66-642

SUBJECT AREA Reading NO. OF PUPILS TESTED 461

CHANGE IN MEAN TEST SCORES



PERCENT OF PUPILS BELOW 25th PERCENTILE - NATIONAL NORMS

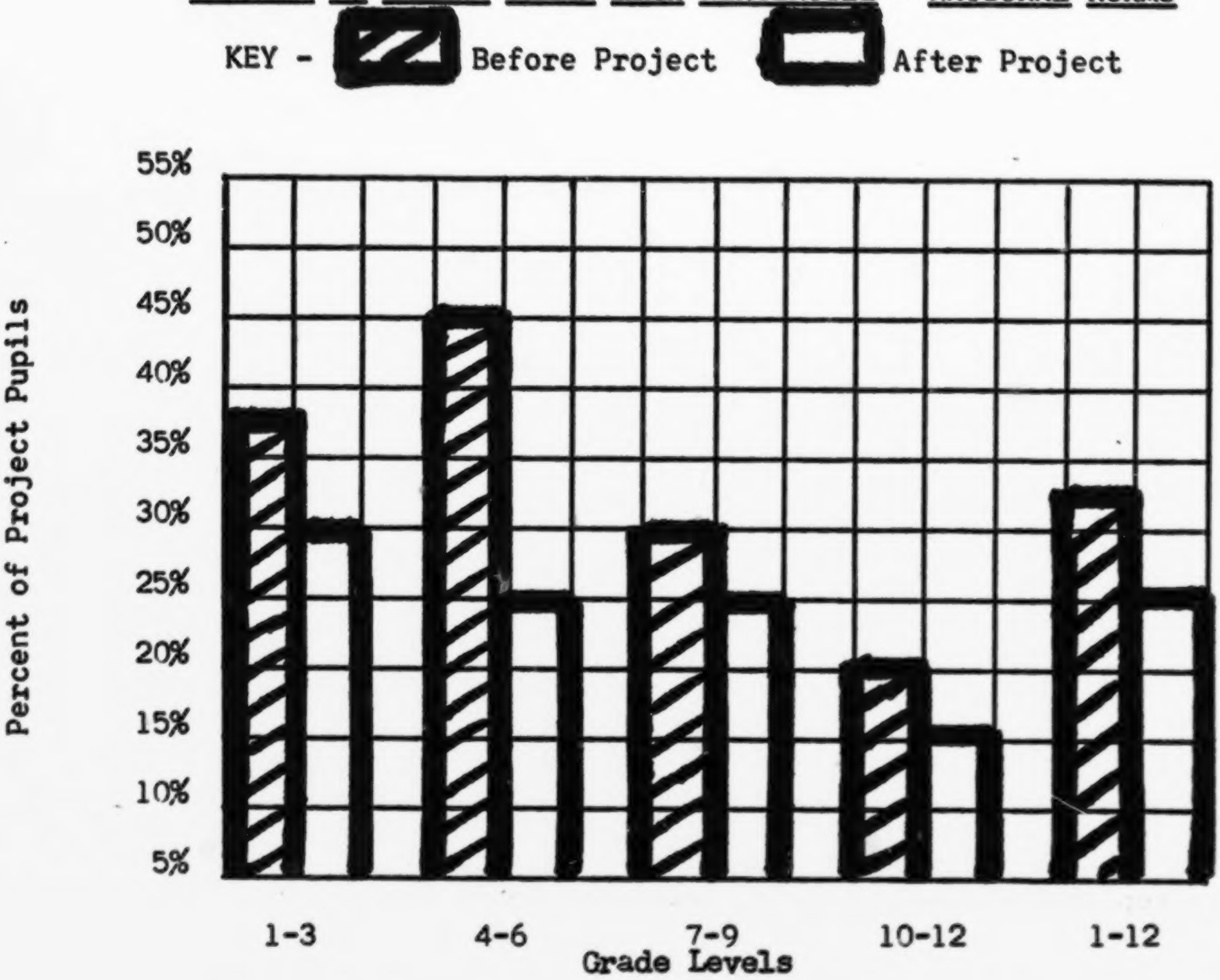


Figure 4. Examples of the use of graphs to reflect change.

Percentages of Pupils Below a National Percentile

A simple and easily understood method of analyzing evaluation data for Title I projects involves counting the number of pupils who score below a designated point or points in a percentile distribution based upon a standardization group. When the group is relatively large, this number can be converted to a percentage. For this type of analysis, it is convenient to use the norms published with the standardized test and to select one or more points for comparison. Comparison of project groups is facilitated when the numbers in the groups being compared are nearly the same, or when both are above 100 in size. The reason for requiring that the numbers on which percentages are based be comparable or large is that percentages based upon very small numbers are easily misinterpreted. For example, 6 cases become 75% when the number in a group contains only 8 pupils, but 6 cases become 2% when the group contains 300 pupils.

The number of cases falling at or below some specified point on national norms reported in percentiles can be obtained either by counting the number of pupils in the group concerned who obtain the raw score corresponding to the selected percentile point or, if the conversions to percentiles have been made for each individual pupil, by simply counting the number of pupils below percentile point 25 or below the 24th percentile rank. In most instances, the 25th percentile can be used satisfactorily for the comparison point. However, the 10th or 15th percentiles may be more appropriate for an extremely disadvantaged group. An example of the summarization of such an analysis is found in Table 3. These results could also be depicted graphically.

Table 3

Percentages of Pupils Below the 25th and 10th Percentiles on National Norms

Condition	7th gr.		8th gr.		9th gr.	
	Pre	Post	Pre	Post	Pre	Post
Total Number	340	302	290	241	236	205
Project	170	161	145	121	118	104
Control	170	141	145	120	118	101
Percent Below 25th Percentile						
Project	32%	19%	36%	24%	37%	31%
Control	31%	28%	37%	33%	39%	30%
Percent Below 10th Percentile						
Project	14%	11%	16%	12%	18%	17%
Control	13%	13%	19%	17%	20%	16%

Statistical Significance of Differences

The analyses and summarization procedures described in the foregoing sections have been based primarily upon description without attempt to generalize the results. In many cases, local evaluators may wish to determine the statistical significance of their results. Such analyses can be readily accomplished following appropriate procedures described in textbooks on statistical methods. Available computer programs can also be used readily for this purpose.

In testing the statistical significance between differences, several considerations are important. These include: 1) whether the distributions being compared represent correlated or independent samples; 2) whether the distributions can be assumed to be random samples from normally distributed populations; 3) whether the variances of the distributions are comparable and 4) whether the groups being compared to assess change were at similar levels at the start of the project.

Types of Samples. Two types of distributions are encountered in evaluation data analyses. These are referred to as correlated or independent samples. Correlated samples result from testing the same individuals twice or from composing groups by pairing the individuals forming the groups on the basis of some characteristic related to the comparison criterion. Thus, a control group matched man-for-man with a project group on the basis of aptitude will yield correlated distributions when the two groups are given achievement tests. In other words for each member of the project group there is a corresponding member of the control group. In the case of pupils being tested twice, each pupil has a score in each distribution.

Independent samples result from testing two separate and unrelated groups. These are assumed to be comparable prior to the project but are not related on an individual or man-for-man basis.

Evaluators computing tests for significance must be especially cognizant of samples, for conclusions may differ if the analysis is computed using the inappropriate method.

Random Sampling. Since the purpose of tests of statistical significance is that of drawing conclusions which can be generalized to groups other than those studied, care must be exercised to make sure that the assumption of random sampling from a normally distributed population is a reasonable one. In most Title I projects, the population to which generalization is to be made will be a relatively unique one and the evaluator cannot follow rigorous sampling procedures. In many instances, however, the assumption that the sample is one which could be random is reasonable. Preparing a frequency polygon or histogram may be helpful in studying the shape of a distribution, but random sampling must be assumed rather than demonstrated in most instances.

Comparability of Variance. The variance of a distribution is the square of the standard deviation. If the variances within groups being compared differ widely, methods taking this into account should be used in the computations for significance of differences. Tests for homogeneity of variance are described in most textbooks and appropriate formulas for varying conditions will also be indicated. Of the assumptions made for the comparison of groups, violation of the assumption of comparable variance is least serious.

Similar Initial Levels. Because of the possibility of unequal units throughout a scale, sound conclusions can be reached only if the groups being compared have been approximately equal at the start of the project. If the groups being compared are 1) random samples, 2) relatively large, or 3) have been matched on pre test scores, this assumption creates no problem. It is appropriate, however, to test the significance of the difference between groups at the start of any project involving group comparisons to determine their initial similarity. If the groups are shown to be comparable at the start of the project, the use of change as a criterion becomes much more defensible than it would otherwise be.

Correlated Samples

In Table 4 data are shown for 20 pupils participating in a project. The scores are raw scores on a locally-constructed test of ability to apply principles. Since these two distributions represent scores on the same pupils tested twice, the data are correlated samples.

Since the obtained value of _t_ is larger than that required for significance, it is concluded that the pupils in this project have made a statistically significant growth on this test. Statistical significance is here defined as evidence that the chances of the two distributions being drawn from one population as a result of sampling fluctuation are less than 5 in 100.

Independent Samples

When the evaluation data to be analyzed are based on independent samples, the statistical analysis is modified to take into account the lack of correlation between the two distributions but the rationale is the same.

Table 4

Raw Scores for Twenty Project Pupils on a Pre and Post Test of Ability to Apply Principles

Pupil Number	Initial Test X_1	Final Test X_2	Difference D $(X_2 - X_1)$	Difference Squared D^2
1	36	37	1	1
2	16	18	2	4
3	19	46	27	729
4	25	18	-7	49
5	34	53	19	361
6	30	26	-4	16
7	29	28	-1	1
8	5	33	28	784
9	29	30	1	1
10	18	20	2	4
11	30	35	5	25
12	15	25	10	100
13	19	19	0	0
14	12	18	6	36
15	16	14	-2	4
16	30	40	10	100
17	19	19	0	0
18	35	37	2	4
19	21	21	0	0
20	16	21	5	25
Sum	454	558	104	2,244
Mean	22.7	27.9		

$$\underline{t} = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\dfrac{\Sigma D^2 - \dfrac{(\Sigma D)^2}{N}}{N\ (N-1)}}} = \frac{27.9 - 22.7}{\sqrt{\dfrac{2,244 - \dfrac{(104)^2}{20}}{20\ (19)}}} = 2.46^*$$

$\underline{t}_{05}$ = 2.093 (from a Table of $\underline{t}$-values)

*Significant beyond the .05 level

An example of an analysis involving two independent samples is shown in Table 5.

Analysis of Covariance

Evaluators competent in using advanced statistical techniques may wish to consider some of the more refined statistical analyses such as the analysis of covariance for analyzing their evaluation data at the local level. This procedure permits the control of individual differences in variables related to the criterion of effectiveness without resorting to the laborious matching process.

The Coefficient of Correlation

In many evaluation situations a measure of the extent to which amounts of one variable are associated with amounts of another variable may be derived. The coefficient of correlation has been developed to express such relationships, since it is an indication of association between two variables.

The coefficient of correlation, designated as r, can vary in magnitude from -1.00 to +1.00. The sign indicates the direction of the relationship and the magnitude of the coefficient indicates the degree of association. A coefficient of zero indicates a complete lack of relationship. In Table 5 are shown scores from two administrations of a test to a class. The coefficient of correlation (Pearson Product Moment Coefficient of Correlation) has been computed for these data and is shown below the table.

In computing the coefficient of correlation, it is assumed that the relationship between the two variables is linear (characterized by a straight line). If this assumption is not met, an erroneously low coefficient will be obtained. It is also assumed that the two distributions are such that

Table 5

Scores on a Socialization Scale for an Experimental and a Control Group

(Pupils were randomly assigned to the two groups at the start of the project).

Experimental Group X_1	Control Group X_2
9	5
10	5
10	5
12	6
13	6
13	7
13	8
14	8
14	8
14	10
14	10
14	10
15	10
15	12
17	12
17	13
18	14
19	16
20	17
	20

$N_1 = 19$ $\Sigma X_1 = 271$ $N_2 = 20$ $\Sigma X_2 = 202$

$\Sigma X_1^2 = 4025$ $\Sigma X_2^2 = 2386$

$\overline{X}_1 = 14.26$ $\overline{X}_2 = 10.10$

$$\underline{t} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\Sigma X_1^2 - \dfrac{(\Sigma X_1)^2}{N_1}}{N_1(N_1 - 1)} + \dfrac{\Sigma X_2^2 - \dfrac{(\Sigma X_2)^2}{N_2}}{N_2(N_2 - 1)}}} = \frac{14.26 - 10.10}{\sqrt{.4669 + .9100}} = \frac{4.16}{1.17} = 3.55^{**}$$

$\underline{t}_{01} = 2.870$ (from a Table of $\underline{t}$-values)

**Significant beyond the .01 level

Table 6

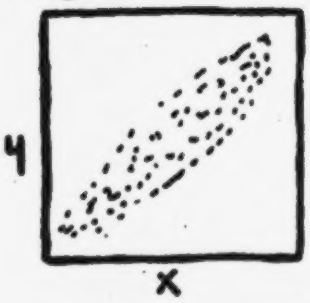
Pearson Product Moment Correlation Worksheet for Scores on Two Administrations of a Test

Pupil Number	1st Admin X	2nd Admin Y	X^2	Y^2	XY
1	27	28	729	784	756
2	27	26	729	676	702
3	30	32	900	1024	960
4	31	31	961	961	961
5	35	37	1225	1369	1295
6	35	36	1225	1296	1260
7	37	39	1369	1521	1443
8	38	37	1444	1369	1406
9	40	42	1600	1764	1680
10	40	40	1600	1600	1600
TOTALS	340	348	11,782	12,364	12,063

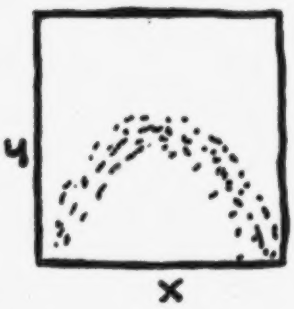
$$r = \frac{\Sigma XY - \frac{\Sigma X \, \Sigma Y}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}} = \frac{12,063 - \frac{(340)(348)}{10}}{\sqrt{\left[11,782 - \frac{(340)^2}{10}\right]\left[12,364 - \frac{(348)}{10}\right]}}$$

$$= \frac{231}{\sqrt{[222]\,[254]}} = .97$$

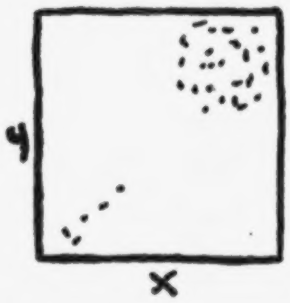
relatively consistent variability exists throughout the range of the plotted values when they are represented as a scattergram. This characteristic is referred to as homoscedasticity. Data which violate the assumption of homoscedasticity yield spuriously high coefficients. Linear and curvilinear relationships, as well as homoscedasticity are illustrated in the following diagrams.



Linear Relationship
with Consistent
Variability



Curvilinear
Relationship



Lack of
Homoscedasticity

An estimate of the Pearson Product Moment coefficient of correlation can be obtained from data expressed as ranks rather than scores. Such an estimate is the Spearman Rank Order Coefficient of Correlation, designated as ρ (lower case Greek rho). In Table 7 the data from Table 6 have been converted to ranks to illustrate the computation of the Spearman Rank Order Coefficient of Correlation. For computational purposes, all values which tie for a given rank are assigned the mean of the rank which would be occupied by the values if no ties existed. Thus, if three individuals have scores of 37 and this is the highest score, all persons would be assigned a rank of two. The next rank assigned would then be four.

Although the Spearman Rank Order Coefficient of Correlation is only an estimate, it is especially useful for expressing relationships when small numbers of cases are involved.

Table 7

Spearman Rank Order Correlation Worksheet for Scores on Two Administrations of a Test

Pupil Number	1st Admin Score	2nd Admin Score	Rank 1st Ad	Rank 2nd Ad	Difference D	D^2
1	27	28	9.5*	9	.5	.25
2	27	26	9.5	10	-.5	.25
3	30	32	8	7	1.0	1.00
4	31	31	7	8	-1.0	1.00
5	35	37	5.5	4.5	1.0	1.00
6	35	36	5.5	6	-.5	.25
7	37	39	4	3	1.0	1.00
8	38	37	3	4.5	-1.5	2.25
9	40	42	1.5	1	.5	.25
10	40	40	1.5	2	-.5	.25
	Total					7.50

*ranks are averaged in case of ties

$$\rho = 1- \frac{6\ \ D^2}{N(N^2-1)} = 1- \frac{6(7.50)}{10(99)} = 1-.045$$

$$= .96$$

Final Analysis

Throughout this section the implication has been that local evaluators may wish to consider a wide variety of evaluation data from Title I projects. Many evaluations will more than satisfy the requirements of State and Federal reporting but by fully evaluating the activities, local, State, and Federal offices will obtain answers to many unique local problems. A thorough analysis will contribute substantially to communicating conclusions, applying findings to new settings, and advancing education.

Pitfalls in Evaluation

As in any professional endeavor, there are many practices and situations in the evaluation process which experience dictates should be avoided. Although, theoretically, an unlimited number of such pitfalls could be listed, the following may have specific implications for evaluation procedures used in Title I projects:

1. Failure to use a sufficiently sensitive instrument to reflect change. Any pupil in a school situation has an almost infinite number of educational stimuli influencing his behavior. It is only realistic to assume, therefore, that changing only a limited number of these influences through a Title I project will result in a small change in behavior. This is especially true if the experimental experience lasts a relatively short time, for example, less than a year. It is therefore very important to use evaluation instruments whose units of measurement are sufficiently sensitive and reflect small increments of change in behavior. Just as it would be unrealistic to weigh diamonds on a cattle scale, it would be also unrealistic to attempt to measure change with an insensitive instrument.

2. Too short a project period for change in behavior to occur. Closely related to lack of sensitivity in a measuring instrument is an unrealistically short project period. Many people, even young children, have spent years developing the habit patterns they now exhibit. To change these behaviors by spending a few hours a day with them for a few weeks is often an overexpectation. Obviously, the larger the segment of educational endeavor involved in a Title I project, the shorter the time required for change to occur. Many projects will have an impact, in some cases, a major one, on

only a small aspect of a pupil's behavior. Therefore, a relatively long period of time will lapse before change occurs.

3. Failure to obtain baseline data or initial measurements. Throughout this document, the importance of pre-test data has been emphasized. When this emphasis is translated into practice, this pitfall can be avoided. Use of the measuring instruments normally administered in the established testing program of the school system is helpful in avoiding this pitfall because the pre-test data are already available. This, of course, assumes that the tests already in use are adequate for the evaluation of project objectives. When regular testing programs are not sufficient to measure all objectives, new instruments should be selected or constructed during the planning stage.

4. Tendency to state objectives in terms of available measuring instruments. Since measures of educational achievement are readily accessible, there is a tendency to make project objectives conform to available instruments rather than to state the objectives as natural outgrowths of a need and then select instruments to measure the objectives. This tendency can be overcome by adhering closely to the steps recommended for developing and planning the evaluation of a project. Even though no measuring devices may be apparent at the time a project objective is formulated, the objective should nevertheless be stated and an intensive effort should be made to evaluate it.

5. Failure to avoid the influence of biased ratings. When project pupils are singled out by their teachers or other educators for special ratings of a subjective nature, there may be a subconscious tendency to

rate the project pupils unduly high either to please the project director or to ensure that the project outcomes will be in a socially desirable direction. Procedures designed to avoid this pitfall include: 1) rating project and non-project pupils without identification; 2) training teachers to be objective in their ratings; 3) utilizing raters or observers who are not identified with the project and therefore have no vested interest; 4) assessing the accuracy of ratings through checks for internal consistency; 5) utilizing several independent raters; and 6) employing several related rating scales in order to determine consistency of ratings.

6. Failure to consider the analysis of evaluation data in the planning stages of a project. Unless the analyses of evaluation data are identified during the planning stages of a project, inefficiency may be encountered in at least two areas: appropriateness of the data collected and ease in quantifying observations. Without adequate advance planning, much hit-and-miss collection of data may occur. In addition, it is appropriate to establish codes, weighting systems, and systems for quantifying observations prior to the actual collection of the data.

7. Failure to involve participating professionals in planning. Some authors of good proposals have failed to communicate effectively with teachers and others responsible for carrying out Title I activities. When representatives of the participants do not engage in the preparation of the proposal, some real needs that could be included are overlooked. As a result of not fully understanding the goals, activities often receive an improper emphasis. When the proposal is written by one person, thorough briefing and continued supervision is important. Those persons skilled

in proposal writing function most efficiently when they work cooperatively during all stages of Title I programing from the development of basic ideas to preparation of the final report.

8. Failure to consider the impact of an educational experience on broad educational outcomes. The importance of evaluating progress toward the objectives of a particular Title I project has been emphasized throughout this document. Indeed, this has been the central focus of the entire discussion. It would be inappropriate, however, to conclude these remarks without encouraging evaluators to look beyond their immediate evaluation data and seek relationships between the project outcomes and the broad educational outcomes postulated for the overall educational process. Such a procedure involves not only the availability of evaluation data from other educational endeavors than the project itself, but imaginative and creative thinking by educators to identify subtle influences and relationships. When these relationships are known, however, the total educational process will be better understood.

As experience with evaluation of Title I projects is gained, additional pitfalls may be identified. As these are reported and disseminated, the evaluation of projects will be improved and the contribution of Title I to increasing the educational attainments of the disadvantaged will be enhanced.

Preparation of a Final Report

At the end of each project, a final summary of the work must be prepared and submitted to the State Educational Agency. At the State level, the reports must be summarized and forwarded to the U. S. Office of Education. It is quite likely that the required data and narrative material will vary through the years, as the needs at each level vary according to the ability to evaluate the worth of Title I projects. Each State will prepare guidelines to facilitate this reporting task, which should be viewed as an opportunity to see how much change has been effected so that efficient plans for next year can be developed both locally and elsewhere in similar situations. Only an objective and honest appraisal of the project can provide the right information for the improvement of American education.

Although the requirements vary from State to State, parts of the requests for evaluation information from the Connecticut and Ohio Departments of Education provide examples of desirable content. The following outline was issued by Connecticut:

I. Description of Project Group
  A. Number of pupils (if sampling or subgrouping are used in evaluation, please specify)
  B. Age Range
  C. Grade Level(s)
  D. Bases for inclusion in project

II. Brief Restatement of Project Objectives and Desired Learning Outcomes

III. Program of Evaluation
  A. Sequence
    1. Base-line data-what outcomes and when measured
    2. On-going measurements-what outcomes, when measured
    3. Final data
  B. Procedures employed
    1. If standardized tests, give title, form, level, publisher
    2. If techniques and instruments were developed specifically

for this project, please describe and include sample copies

C. Pupil Characteristics and Behaviors Evaluated

IV. Results - Data Presentation and Analysis.

V. Overall Evaluation of Project and Recommendations

Four of the forms used in Ohio are particularly worthy of reproduction because of the references to objectives and the concise manner in which data can be reported. The forms appear here as Figures 5, 6, 7, and 8.

In addition to these and other data, the following outline was used to obtain narrative information in Ohio:

Title I Project Narrative Evaluation Report
Part III

Directions: This section is to be completed for each project. Complete a response for each question. Be concise.

1. (a) Appraise this project in relation to the general improvement of educational opportunities provided the educationally deprived youth in your school district.

   (b) Report either positive or negative outcomes of the project which were not anticipated. Reference is to be made to those outcomes which relate to stated objectives as well as those unrelated to the stated objectives.

2. (a) Report summarized statistical or subjective data which show that the project substantially changed the achievement level of project participants.

   (b) What has been conclued as a result of thses achievement data? What do these data mean?

3. (a) Report summarized statistical or subjective data which shows that the project substantially changed the behavior, attitudes, or self concept of project participants.

   (b) What has been concluded as a result of these behavioral, attitudinal or self conceptual data? What do these data mean?

EVALUATION OF PROJECT OBJECTIVES

(This page is to be completed only for those objectives which were evaluated by objective instruments or devices Standardized on state and/or national levels.)

Number of students in project ________

Average number of hours each child was involved in the project ________

Rank listing of the 4 most important objectives of this project. Categorize according to appropriate coded listing on reverse side List below only the code number.	Tests or measurement devices used to evaluate each objective. Record below only the code number obtained from listing on reverse side.	Units in which data reported. Record below only the code number obtained from the listing on reverse side.	Expected Change Score: The amount of change expected for this group considering the duration of the project and the ability of the group.	Median Baseline Data: The midpoint on a Scale of a frequency distribution of the project group prior to or at the beginning of the project.	Median Terminal Data: the mid point on a Scale of a frequency distribution of the project group at the end of the project.	Median Difference Score Formula: median terminal data minus median baseline data.	Median Difference Shows: (check one)				Additional Statistical and/or Anecdotal Data Which Expands Qualifies, or Justifies Your Judgment About the Difference Score.
							Marked Improvement	Improvement	No Improvement	Decrease	
1.											
2.											
3.											
4.											

Figure 5. Form used in Ohio to summarize the evaluation of project objectives measured by objective data.

CODE FOR OBJECTIVES

Objectives

- To increase general achievement
- To increase school readiness
- To increase reading skills in general
- To increase reading vocabulary skills
- To increase reading comprehension skills
- To increase arithmetic comprehension
- To improve language arts and/or communication skills
- To increase understanding and knowledge of science
- To increase facility with and knowledge of industrial art
- To expand understandings of social sciences
- To increase an awareness and an appreciation of the humanities - art, music, literature, cultural development
- To acquaint students with library services and/or materials resource centers
- To overcome speech defects
- To improve business education skills
- To improve study skills
- To improve physical development through physical education and recreation
- To improve physical health through medical and/or dental treatment
- To improve nutrition
- To improve school attendance
- To reduce school droupout rate
- To improve self concept
- To improve attitudes and increase interests toward school-type activites
- To improve emotional health
- Other, specify ________________

____

____

____

____

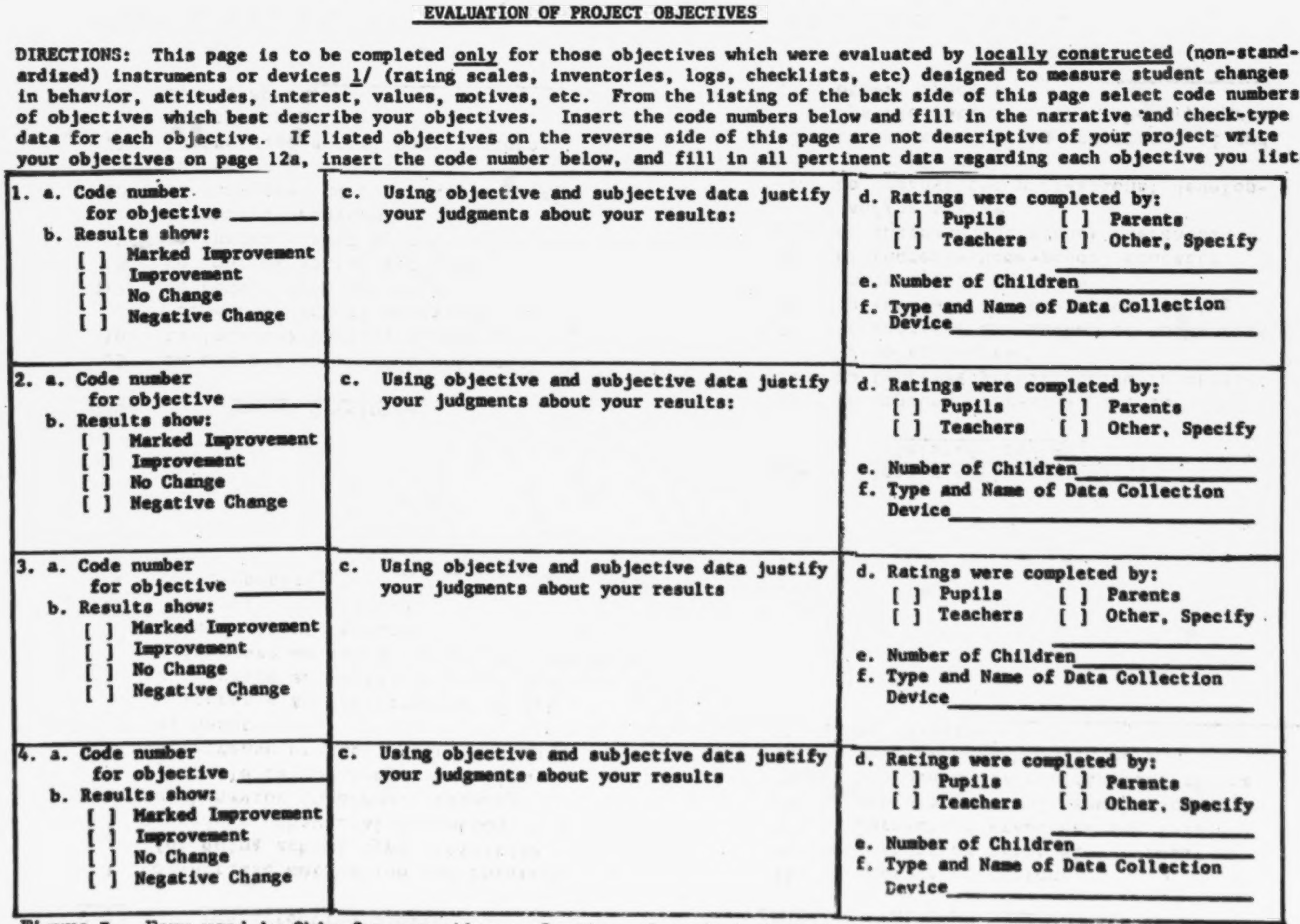
CODE FOR TESTS, DEVICES, INSTRUMENTS

Code Number	Tests, Devices
41	Lee-Clark Readiness
42	Metropolitan Readiness
43	Iowa Test of Basic Skills
44	Stanford Achievement
45	California Achievement
46	Metropolitan Achievement
47	Wide Range Achievement Test
48	Monroe Reading
49	Durrell-Sullivan Reading Capacity
50	Durrell Analysis of Reading Difficulties
51	Botel Reading Inventory
52	Ohio School Survey
53	SRA Achievement Series
54	Davis Reading Tests
55	Durost Work Mastery Test
56	Gates Basic Reading
57	Gates Primary Reading Tests
58	Gillmore Oral Reading Test
59	Gray Oral Reading Test
60	Iowa Silent Reading Test
61	Kelley-Greene Reading Comprehension
62	Nelson Reading Test
63	California Test of Mental Maturity
64	Chicago Non-Verbal
65	Henmon Nelson Test of Mental Ability
66	Lorge-Thorndike Intelligence
67	Otis Quick Scoring Mental Ability Test
68	SRA Primary Mental Abilities
69	SRA Test of General Ability
70	Stanford Binet Intelligence Scale
71	Wechsler Intelligence Scale for Children
72	Pupil Self Rating Scale, specify ____________
73	Teacher Rating Scale, specify
74	Parent inventory, specify ____
75	Self Concept Inventories, specify ____________
76	Other, specify ________ ________________ ________________

CODE FOR UNITS WHICH DATA REPORTED

Code Number	Units
91	Total Grade Equivalent Score
92	Modal Grade Equivalent Score
93	IQ
94	Percentiles
95	Standard scores
96	Scholastic marks
97	Days
98	Other, specify____ ____________
99	____________ ____________
100	____________ ____________
101	____________ ____________
102	____________ ____________
103	____________ ____________

Figure 6. Codes used to complete the form in Figure 5.



EVALUATION OF PROJECT OBJECTIVES

DIRECTIONS: This page is to be completed only for those objectives which were evaluated by locally constructed (non-standardized) instruments or devices 1/ (rating scales, inventories, logs, checklists, etc) designed to measure student changes in behavior, attitudes, interest, values, motives, etc. From the listing of the back side of this page select code numbers of objectives which best describe your objectives. Insert the code numbers below and fill in the narrative and check-type data for each objective. If listed objectives on the reverse side of this page are not descriptive of your project write your objectives on page 12a, insert the code number below, and fill in all pertinent data regarding each objective you list.

1. a. Code number for objective ____ b. Results show: ☐ Marked Improvement ☐ Improvement ☐ No Change ☐ Negative Change	c. Using objective and subjective data justify your judgments about your results:	d. Ratings were completed by: ☐ Pupils ☐ Parents ☐ Teachers ☐ Other, Specify ____ e. Number of Children ____ f. Type and Name of Data Collection Device ____
2. a. Code number for objective ____ b. Results show: ☐ Marked Improvement ☐ Improvement ☐ No Change ☐ Negative Change	c. Using objective and subjective data justify your judgments about your results:	d. Ratings were completed by: ☐ Pupils ☐ Parents ☐ Teachers ☐ Other, Specify ____ e. Number of Children ____ f. Type and Name of Data Collection Device ____
3. a. Code number for objective ____ b. Results show: ☐ Marked Improvement ☐ Improvement ☐ No Change ☐ Negative Change	c. Using objective and subjective data justify your judgments about your results	d. Ratings were completed by: ☐ Pupils ☐ Parents ☐ Teachers ☐ Other, Specify ____ e. Number of Children ____ f. Type and Name of Data Collection Device ____
4. a. Code number for objective ____ b. Results show: ☐ Marked Improvement ☐ Improvement ☐ No Change ☐ Negative Change	c. Using objective and subjective data justify your judgments about your results	d. Ratings were completed by: ☐ Pupils ☐ Parents ☐ Teachers ☐ Other, Specify ____ e. Number of Children ____ f. Type and Name of Data Collection Device ____

Figure 7. Form used in Ohio for reporting evaluation of project objectives measured by non-standardized instruments.

Code	The Affective Domain
1	to increase motivation and interest for doing school type activities
2	to reduce behavioral deviation, i.e., misbehavior in school, truancy juvenile delinquency, vandalism
3	to increase preception and awareness of beauty
4	to increase an appreciation of art
5	to increase an appreciation of music
6	to increase an appreciation of literature
7	to improve self concept
8	to stimulate curiosity
9	other, specify________________________

Code	Physical Development
25	to improve nutrition
26	to correct or treat vision loss
27	to correct or treat hearing loss
28	to improve physical growth patterns
29	to improve motor coordination skills
30	to improve speech patterns
31	other, specify________________________

Code	Social Development
15	to improve attendance patterns
16	to increase participation in organized school activites
17	to reduce drop out rate
18	to improve social attitude
19	to increase self referrals to counselors/psychologists
20	to increase independent(positive) behavior
21	to increase social environmental awareness
22	other, specify__________________

Code	Academically and School Related Problems
35	to improve work-study skills
36	to increase participation in classroom activites
37	to increase the number of books read
38	to increase the amount of reading material in the home
39	to increase home-school contacts
40	to increase vocational awareness and/or skills
41	to improve the professionsl development of teachers
42	to increase parent acceptance of and participation in the school program
43	other, specify __________________

Figure 8. Codes used to record the more frequently used objectives for Figure 6.

4. (a) Report the degree of effectiveness of the procedural phases (the methods by which you attempted to change the achievement, behavior, etc., of children) of the project. Some of your procedures may have been more effective than others. Summarize these differences.

   (b) Report those equipment and/or materials which were of considerable assistance in attaining behavioral change in children.

   (c) Report those equipment and/or materials which were of little or no value in attaining behavioral change in children.

5. (a) Summarize Title I teacher response and reactions to the Title I project.

   (b) Summarize non-Title I teacher (those within and those outside target areas) responses and reactions to the Title I project.

   (c) Summarize administrative (building administrators, supervisors, and central office administration) responses and reactions to the Title I project.

   (d) Summarize community reactions to this Title I project.

6. (a) On the basis of all the above comments (items 1-5) report how they will affect your plans for future Title I projects.

   (b) How will these data (items 1-5) influence the existing curriculum (regular non-Title I educational programs) in your school district?

Note: We would appreciate your attaching to this report sample copies of locally developed instructional materials or guides which were devised specifically for the educationally deprived.

The order in which information is reported is not as important as the content. The evaluation should convey to the reader the methods employed to improve rate of learning in specified children and the degree to which these methods were successful. If significant differences are not observed, disappointment should not reign. Non-significant differences are often found when innovative action is evaluated. A good explanation of the findings is often at hand. The findings must be reported, regardless of the differences noted, so that evaluators in other settings will know the effect of the various Title I activities as they are applied and evaluated throughout the nation.

Summary Statement

Reference has been made repeatedly throughout this document to the emphasis which evaluation of educational outcomes will receive under Title I. To the extent that educators successfully meet the challenges posed by the evaluation of new projects and programs under Title I, the basic intent of the legislation will be satisfied. Of even greater significance, however, will be the advances in the education of educationally disadvantaged pupils which will result.

Selected References

Keeping up with current methods, materials, and procedures used in evaluation is a difficult task. To stay in touch with current thinking, evaluators have to read widely in the professional literature and converse with other evaluators. A number of references are available which contain expanded discussions of many of the points in this document. The references noted on the following pages do not constitute a comprehensive list, but do suggest that much information about evaluation is available. Furthermore, the new emphasis on evaluating educational change should stimulate many additional, and a few better, volumes.

Many references review, abstract, or describe journal articles, published texts, or other books. As such, they are secondary, rather than primary, sources of information and are useful in providing general and sometimes specific, but translated, information and knowledge about evaluation. They can guide the reader to the sources of much relevant information.

Although reference works are issued or re-issued periodically, the publication lag keeps them from being up to date. First, there is a lag between data collection and interpretation and publication. Another lag occurs in the assimilation of information and the publication of reference works. Seldom do reference works carry information that is less than a year old; most of the "current" information in references is three to five years old.

Reference Volumes

The reference volumes cited here are of three types: 1) test descriptions and reviews; 2) abstracts and indexes of measurement literature; 3) surveys and analyses of educational research.

Tests in Print and The Mental Measurements Yearbook are the most comprehensive and current bibliographic references describing standardized tests. (Earlier bibliographies, primarily useful for those who need historical information about tests, are not described here.)

Child Development Abstracts and Bibliography, Dissertation Abstracts, Education Index, and the Psychological Abstracts are useful tools for locating evaluation literature on education published as books, journal articles, or doctoral dissertations.

The Annual Review of Psychology, the Review of Educational Research, and the Encyclopedia of Educational Research periodically survey and analyze research in psychology and education, including educational measurement.

Buros, O.K. (Ed.) Tests in Print: A Comprehensive Bibliography of Tests for Use in Education, Psychology, and Industry, Highland Park, N.J.: The Gryphon Press, 1961.

Buros, O.K. (Ed.) The Sixth Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 1965. (Earlier Yearbooks published by Rutgers University Press and The Mental Measurements Yearbook. Yearbooks are not published yearly.)

Child Development Abstracts and Bibliography, Chicago, Ill.: University of Chicago Press. (A publication of the Society for Research in Child Development.)

Dissertation Abstracts, Ann Arbor, Mich.: University Microfilms, Inc.

Education Index, New York: H.W. Wilson Company.

Gage, N.L. (Ed.) *Handbook of Research on Teaching*, Chicago: Rand McNally, 1963. (A project of the American Educational Research Association.)

Harris, C. W. (Ed.) *Encyclopedia of Educational Research*, (3rd Ed.) New York: Macmillan, 1960. (A project of the American Educational Research Association.)

Psychological Abstracts, Washington, D.C.: American Psychological Association, Inc.

Books

Books are important sources of information about tests, measurement, statistics, design, and policy. Whereas many books discuss the same or similar topics, the discussions differ in emphasis and in assumption of the reader's prior knowledge. The lists in this section are not intended to be comprehensive, but will provide more than an adequate library for evaluators concerned with Title I projects and procedures.

The pamphlets or books in the following list will be very helpful to most evaluators and can be obtained without charge or for a small fee:

Bernstein, A.L. *A Handbook of Statistical Solutions for the Behavioral Sciences*. New York: Holt, Rinehart and Winston, 1964.

Bloom, B.S. (Ed.) et al. *Taxonomy of Educational Objectives I: Cognitive Domain*. New York: Longmans Green, 1956.

Bradley, J. I. & McLelland, J.N. *Basic Statistical Concepts*. Chicago: Scott Foresman, 1963.

Deutsch, M., Fishman, J.A., Kogan, L., North, R. & Whitman, M. *Guidelines for Testing Minority Group Children*. Journal of Social Issues 22: (Supl) 127-145, 1964. (Available separately from SPSSI, P.O. Box 1248, Ann Arbor, Michigan.)

Diederich, P.B. *Short-cut Statistics for Teacher Made Tests*. Princeton, N.J.: Educational Testing Service, 1960. (Evaluation and Advisory Service Series, No. 5.)

Elzey, F.F. *A Programmed Introduction to Statistics*. Belmont, Calif.: Wadsworth, 1966.

Katz, M. R. (Ed.) *Locating Information on Educational Measurement: Sources and References*. Princeton, N.J. Educational Testing Service, 1965. (Evaluation and Advisory Service Series, No. 1.)

Katz, M. R. *Selecting an Achievement Test: Principles and Procedures*. Princeton, N. J.: Educational Testing Service, 1958. (Evaluation and Advisory Service Series, No. 3.)

Krathwohl, D.R., Bloom, B.S. and Masia, B.B. *Taxonomy of Educational Objectives II: Affective Domain*. New York: David McKay, 1964.

Lyman, H. B. *Test Scores and What They Mean*. Englewood Cliffs, N.J.: Prentice Hall, 1963.

McCollough, C. & Van Alta, L. *Statistical Concepts: A Program for Self Instruction*. New York: McGraw Hill, 1963.

McLaughlin, K.F. (Ed.) *Understanding Testing*. Washington, D.C.: U.S. Government Printing Office, 1962. (U.S. Office of Education, OE-25003.)

Neidt, C.O., Ivanoff, J. and Peterson, F. *Workbook for Statistical Methods in Educational and Psychological Research*. Dubuque, Iowa: Wm. Brown, 1965.

Schoer, L. *An Introduction to Statistics and Measurement*, Boston, Allyn and Bacon, 1966.

Information pertaining to the content of each book and the audience to whom it is directed for the books on the longer (and more expensive) list which is below can be obtained from the current catalogs from each book publisher. Reviews of books can be found in *Contemporary Psychology* or *Educational and Psychological Measurement* and reviews of many of them can be found in other professional journals. Most of the volumes have been annotated in the pamphlet by Katz cited above. Most educational or psychology libraries in colleges and universities have single copies that could be reviewed prior to a decision to purchase any volume. Although the titles of most books explain the content, a brief classification note will follow most references.

Ahmann, J.S. & Glock, M.D. *Evaluating Pupil Growth*. Boston: Allyn and Bacon, 1963. Measurement in education.

Anastasi, Anne. *Psychological Testing*. (2nd ed.) New York: Macmillan, 1961. Measurement theory and methods.

Bauernfeind, R.H. *Building a School Testing Program*. Boston: Houghton Mifflin, 1963. Measurement in education.

Chauncey, H. & Dobbin, J. Testing: *Its Place in Education Today*. 1st. ed. New York: Harper & Row, 1963. Commentary on testing.

Cronbach, L.J. *Essentials of Psychological Testing*. (2nd ed.) New York: Harper & Brothers, 1960. Measurement theory and methods.

Davis, F. B. *Educational Measurements and Their Interpretation*. Belmont, Calif.: Wadsworth, 1964. Measurement theory and methods.

Downie, N.M. & Heath, R.W. *Basic Statistical Methods*. New York: Harper and Row, 1965. Basic statistics.

Durost, W.N. & Prescott, G.A. *Essentials of Measurement for Teachers*. New York: Harcourt, Brace, and World, 1962. Measurement in education.

Ebel, R.L. *Measuring Educational Achievement*. Englewood Cliffs, N.J.: Prentice Hall, 1965. Measurement theory and methods.

Edwards, A. L. *Statistical Methods for the Behavioral Sciences*. New York: Rinehart, 1964. Statistics.

Findley, W.G., (Ed.) *The Impact and Improvement of School Testing Programs*. 62nd Yearbook, Part II. Chicago: National Society for the Study of Education, 1963.

Freeman, F.S. *Theory and Practice of Psychological Testing*. (3rd ed.) New York: Holt, Rinehart and Winston, 1962. Measurement theory and methods.

Froehlich, C. P. & Hoyt, K.B. *Guidance Testing and Other Student Appraisal Procedures for Teachers and Counselors*. (3rd ed.) Chicago: Science Research Associates, Inc., 1959.

Good, C.V. *Introduction to Educational Research*. New York: Appleton-Century Crofts, 1963.

Goslin, D.A. *The Search for Ability*: *Standardized Testing in Social Perspective*. New York: Russell Sage, 1963. Commentary on testing.

Green, J. A. *Teacher Made Tests*. New York: Harper, 1963.

Guba, E. (Ed.) *The Training and Nurture of Educational Researchers*. Bloomington, Indiana: Phi Delta Kappa, 1965.

Guilford, J. P. *Psychometric Methods*. (2nd ed.) New York: McGraw-Hill, 1954. Advanced measurement methods.

Helmstadter, G. C. *Principles of Psychological Measurement*. New York: Appleton-Century Crofts, 1964. Measurement theory and methods.

Kerlinger, F.N. *Foundations of Behavioral Research*. New York: Holt, Rinehart and Winston, 1964.

Lavin, D.E. *The Prediction of Academic Performance*. New York: Russell Sage, 1965. Commentary on testing.

Lindquist, E.F. *Design and Analysis of Experiments in Psychology and Education.* Boston: Houghton-Mifflin, 1953. Advanced methods in planning experiments.

Miller, D.C. *Handbook of Research Design and Social Measurement*. New York: David McKay, 1964.

Passow, A.H. (Ed.) *Nurturing Individual Potential*. Washington, D.C.: National Education Association, 1964.

Rupiper, O.J. Item Writing: *A Programed Test of Rules for Writing Objective Type Items*, Norman, Okla.: Harlow, 1964.

Selltiz, C., Jahoda, Marie, Deutsch, M., and Cook, S.W. *Research Methods in Social Relations*. New York: Holt, Rinehart and Winston, 1962.

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill, 1956. Statistics.

Stanley, J.C. *Measurement in Today's Schools*. (4th ed.) Englewood-Cliffs, N.J.: Prentice-Hall, 1964. Measurement in education.

Super, D.E. & Crites, J.O. *Appraising Vocational Fitness by Means of Psychological Tests*. (Rev.) New York: Harper, 1962.

Soloman, H. (Ed.) *Studies in Item Analysis and Prediction*. Stanford, Calif.: Stanford University Press, 1961.

Travers, R. N. W. *An Introduction to Educational Research*. New York: Macmillan, 1964.

Tyler, Leona E. *Tests and Measurements*. Foundations of Modern Psychology Series. Englewood Cliffs, N.J.: Prentice-Hall, 1963. Measurement in education.

Wert, J.W., Neidt, C.E. & Ahman, J.S. *Statistical Methods in Educational and Psychological Research*. New York: Appleton-Century-Crofts, 1954. Statistics.

Professional Journals

It is easier to keep up to date by reviewing pertinent journals than through other reference works. Journal articles have a further advantage in providing a first hand reference to specific studies as well as secondary source material found when the author traces the history of a problem and discusses the results of his study in connection with the findings of other researchers.

Information applicable to evaluation appears in professional journals in a variety of ways: reports of experiments, surveys of evaluation literature, surveys of procedures followed, studies of evaluative techniques, applications of evaluative techniques in many areas, and reviews of books on evaluation activities. Since each journal limits its coverage to certain areas, review of several journals is necessary to stay in tune with current thinking.

Although the following list could be much longer, it contains the journals of major significance to evaluators. These journals vary greatly in their demands on the reader.

American Educational Research Journal. American Educational Research Association. Quarterly.

Educational and Psychological Measurement. Box 6907, College Station, Durham, North Carolina. Quarterly.

Journal of Educational Measurement. National Council on Measurement in Education. Semi-annual.

Journal of Educational Psychology. American Psychological Association, Inc. Bimonthly.

Journal of Educational Research. Dembar Educational Services, Inc., Box 1605, Madison, Wisconsin. 10 issues per year.

Journal of Experimental Education. Dembar Educational Services, Inc., Box 1605, Madison, Wisconsin. Quarterly.

The Personnel and Guidance Journal. American Personnel and Guidance Association. Monthly (September through June).

Psychometrika. Psychometric Society. Quarterly.

Test Publishers

Although it is very helpful to read reviews of tests in The Mental Measurement Yearbooks, professional journals, and textbooks, the final selection of a test should take place after a review of a specimen or regular set of the test materials and the accompanying technical manual. Most available tests can be found in Tests in Print, but those published later must be located in the cataglogue of the publisher. Publishers will furnish detailed information about tests upon request, and many publishers will answer questions about their tests.

Catalogs usually provide brief descriptions of each test, information about scoring, and procedures for ordering tests. Publishers furnish, for a small fee in most instances, specimen sets of most tests. The specimen set usually includes a copy of the test, an answer sheet, a manual, and related materials. Review of the test should determine how well the items measure the stated objectives of the program.

Distribution of some tests are restricted to preserve security and to protect them from use by persons who are not adequately prepared to use them. A list of test publishers and their addresses appear below.

American Guidance Service, Inc., 720 Washington Avenue, S.E., Minneapolis, Minnesota 55414.

The Bobbs-Merrill Company, Inc., 4300 West 62nd Street, Indianapolis, Indiana 46268.

Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas 66802.

The Bureau of Educational Research and Service, East Hall, State University of Iowa, Iowa City, Iowa 52240.

Bureau of Publications, Teachers College, Columbia University, New York, New York 10027.

California Test Bureau, Del Monte Research Park, Monterey, California 93940.

The Center For Psychological Service, 1835 "Eye" Street, N.W., Washington, D. C. 20006.

Committee on Diagnostic Reading Tests, Inc., Mountain Home, North Carolina 28758.

Consulting Psychologists Press, Inc., 577 College Avenue, Palo Alto, California 94306.

Cooperative Test Division, Educational Testing Service, Princeton, New Jersey 08540.

Harcourt, Brace and World, Inc., 757 3rd Avenue, New York, New York 10017.

Houghton Mifflin Company, 53 West 43rd Street, New York, New York 10036.

Institute For Personality and Ability Testing, 1602-04 Coronado Drive, Champaign, Illinois 61822.

Personnel Press, Inc., 20 Nassau Street, Princeton, New Jersey 08540.

Personnel Research Institute, Western Reserve University, Cleveland, Ohio 44106.

The Psychological Corporation, 304 East 45th Street, New York, New York 10017.

Psychological Test Specialists, Box 1441, Missoula, Montana 59801.

Psychometric Affiliates, Box 1625, Chicago, Illinois 60690.

Science Research Associates, Inc., 259 East Erie Street, Chicago, Illinois 60611.

Sheridan Supply Company, P.O. Box 837, Beverly Hills, California 90213.

C.H. Stoelting Company, 424 North Homan Avenue, Chicago, Illinois 60624.

Western Psychological Services, 12035 Wilshire Boulevard, Los Angeles, California 90025.

Glossary

Definitions of Commonly Used Evaluation Terms*

Achievement age. The age for which a given achievement test score is the real or estimated average. (Also called educational age or subject age). If the achievement age corresponding to a score of 36 on a reading test is 10 years, 7 months (10-7), this means that pupils 10 years 7 months achieve, on the average, a score of 36 on that test.

Achievement test. A test that measures the extent to which a person has "achieved" something-acquired certain information or mastered certain skills, usually as a result of specific instruction.

Age equivalent. The age for which a given score is the real or estimated average score.

Age norms. Values representing typical or average performance for persons of various age groups.

Alternate-form reliability. The closeness of correspondence, or correlation between results on alternate (i.e. equivalent or parallel) forms of a test; thus, a measure of the extent to which the two forms are consistent or reliable in measuring whatever they do measure, assuming that the examinees themselves do not change in the abilities measured between the two testings. (See RELIABILITY, STANDARD ERROR.)

Arithmetic mean. The sum of a set of scores divided by the number of scores. (Commonly called average, mean.)

Battery. A group of several tests standardized on the same population, so that results on the several tests are comparable. Sometimes loosely applied to any group of tests administered together, even though not standardized on the same subjects.

Class analysis chart. A chart, usually prepared in connection with a battery of achievement tests, that shows the relative performance of members of a class on the several parts of the battery.

---

*Reprinted with permission from "A Glossary of 100 Measurement Terms," Test Service Notebook Number 13, by Roger T. Lennon, Director, Division of Test Research and Service, Harcourt, Brace and World, Inc. Tarrytown, New York.

Correlation. Relationship or "going togetherness" between two scores or measures; tendency of one score to vary concomitantly with the other, as the tendency of students of high IQ to be above average in reading ability. The existence of a strong relationship-i.e., a high correlation between two variables does not necessarily indicate that one has any causal influence on the other.

Criterion. A standard by which a test may be judged or evaluated; a set of scores, ratings, etc., that a test is designed to predict or to correlate with. (See VALIDITY.)

Distribution (frequency distribution). A tabulation of scores from high to low, or low to high, showing the number of individuals that obtain each score or fall in each score interval.

Grade equivalent. The grade level for which a given score is the real or estimated average.

Intelligence quotient (IQ). Orginally, the ratio of a person's mental age to his chronological age $\frac{(MA)}{(CA)}$ or, more precisely, especially for older persons, the ratio of mental age to the mental age normal for chronological age (in both cases multiplied by 100 to eliminate the decimal). More generally, IQ is a measure of brightness that takes into account both score on an intelligence test and age. A deviation IQ is such a measure of brightness, based on the difference or deviation between a person's obtained score and the score that is normal for the person's age.

Item analysis. The process of evaluating single test items by any of several methods. It usually involves determining the difficulty value and the discriminating power of the item, and often its correlation with some criterion.

Kuder-Richardson formula(s). Formulas for estimating the reliability of a test from information about the individual items in the test, or from the mean score, standard deviation, and number of items in the test. Because the Kuder-Richardson formulas permit estimation of reliability from a single administration of a test, without the labor involved in dividing the test into halves, their use has become common in test development. The Kuder-Richardson formulas are not appropriate for estimating the reliability of speeded tests.

Median. The middle score in a distribution; the 50th percentile; the point that divides the group into two equal parts. Half of the group of scores fall below the median and half above it.

Mental age (MA). The age for which a given score on an intelligence test is average or normal. If a score of 55 on an intelligence test corresponds to a mental age of 6 years, 10 months, then 55 is presumably the average score that would be made by an unselected group of children 6 years, 10 months of age.

N. The symbol commonly used to represent the number of cases in a distribution, study, etc.

Normal distribution. A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. Figure 3. shows such a graph of a normal distribution, known as a normal curve or normal probability curve. In a normal distribution, scores or measures are distributed symmetrically about the mean, with as many cases at various distances above the mean as at equal distances and decreasing in frequency the further one departs from the average, according to a precise mathematical equation. The assumption that mental and psychological characteristics are distributed normally has been very useful in much test development work.

Objective test. A test in the scoring of which there is no possibility of difference of opinion among scorers as to whether responses are to be scored right or wrong. It is contrasted with a "subjective" test-e.g., the usual essay examination to which different scorers may assign different scores, ratings, or grades.

Percentile (P). A point (score) in a distribution below which falls the per cent of cases indicated by the given percentile. Thus the 15th percentile denotes the score or point below which 15 per cent of the scores fall. "Percentile" has nothing to do with the per cent of correct answers an examinee has on a test.

Percentile rank. The per cent of scores in a distribution equal to or lower than the score corresponding to the given rank.

Personality test. A test intended to measure one or more of the non-intellective aspects of an individual's mental or psychological make-up. Personality tests include the so-called personality inventories or adjustment inventories...which seek to measure a person's status on such traits as dominance, sociablity, introversion, etc., by means of self-descriptive responses to a series of questions; rating scales... which call for rating, by one's self or another, of the extent to which a subject possess certain characteristics; situation tests in which the individual's behavior in simulated life-like situations is observed by one or more judges, and evaluated with reference to various personality traits; and opinion or attitude inventories. Some writers also classify interest inventories as personality tests.

Practice effect. The influcence of previous experience with a test on a later administration of the same test or similar test; usually, an increase in the score on the second testing, attributed to increased familiarity with the directions, kinds of questions, etc. Practice effect is greatest when the interval between testings is small, when the materials in the two tests are very similar, and when the initial test-taking represents a relatively novel experience for the subjects.

Profile. A graphic representation of the results on several tests, for either an individual or a group, when the results have been expressed in some uniform or comparable terms. This method of presentation permits easy identification of areas of strength or weakness.

Quartile. One of three points that divide the cases in a distribution into four equal groups. The lower quartile, or 25th percentile, sets off the lowest fourth of the group; the middle quartile is the same as the 50th percentile, or median; and the third quartile, or 75th percentile, marks off the highest fourth.

Random sample. A sample of the members of a population drawn in such a way that every member of the population has an equal chance of being included-that is, drawn in a way that precludes the operation of bias or selection. The purpose in using a sample thus free of bias is, of course, that the sample be fairly "representative" of the total population, so that sample findings may be generalized to the population. A great advantage of random samples is that formulas are available for estimating the expected variation of the sample statistics from their true values in the total population; in other words, we know how precise an estimate of the population value is given by a random sample of any given size.

Raw score. The first quantitative result obtained in scoring a test. Usually the number of right answers, number right minus some fraction of wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.

Reliability. The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, relative freedom from errors of measurement. Reliability is usually estimated by some form of *reliability coefficient* or by the *standard error of measurement*.

Representative sample. A sample that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purposes under investigation-e.g., in an achievement test norm sample, proportion of pupils from each state, from various regions, from segregated and non-segregated schools, etc.

Split-half coefficient. A coefficient of reliability obtained by correlating scores on one half of a test with scores on the other half. Generally, but not necessarily, the two halves consist of the odd-numbered and the even-numbered items.

Standard deviation (S.D.). A measure of the variability or dispersion of a set of scores. The more the scores cluster around the mean, the smaller the standard deviation.

Standard Error (S.E.). An estimate of the magnitude of the "error of measurement" in a score-that is, the amount by which an obtained score differs from a hypothetical true score. The standard error is an amount such that in about two-thirds of the cases the obtained score would not differ by more than one standard error from the true score. The probable error (P.E.) of a score is a similar measure except that in about half the cases the obtained score differs from the true score by not more than one probable error. The probable error is equal to about two-thirds of the standard error. The larger the probable or the standard error of a score, the less reliable the measure.

Standard score. A general term referring to any of a variety of "transformed" scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. The simplest type of standard score is that which expresses the deviation of an individual's raw score from the average score of his group in relation to the standard deviation of the scores of the group. Thus:

$$\text{Standard score (Z)} = \frac{\text{raw score(x) - mean (M)}}{\text{standard deviation (S.D.)}}$$

By multiplying this ratio by a suitable constant and by adding or subtracting another constant, standard scores having any desired mean and standard deviation may be obtained. Such standard scores do not affect the relative standing of the individuals in the group nor change the shape of the original distribution.

More complicated types of standard scores may yield distributions differing in shape from the orginal distribution; in fact, they are sometimes used for precisely this purpose.

Standardized test (standard test). A systematic sample of performance obtained under prescribed conditions, scored according to definite rules, and capable of evaluation by reference to normative information. Some writers restrict the term to tests having the above properties, whose items have been experimentally evaluated, and/or for which evidence of validity and reliability are provided.

Stanine. One of the steps in a nine-point scale of normalized standard scores. The stanine (short for standard-nine) scale has values from 1 to 9, with a mean of 5, and a standard deviation of 2.

Survey test. A test that measures general achievement in a given subject or area, usually with the connotation that the test is intended to measure group status, rather than to yield precise measures of individuals.

True score. A score entirely free of errors of measurement. True scores are hypothetical values never obtained by testing, which always involves some measurement error. A true score is sometimes defined as the average score of an infinite series of measurements with the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the testings.

Validity. The extent to which a test does the job for which it is used. Validity, thus defined, has different connotations for various kinds of tests and, accordingly, different kinds of validity evidence are appropriate for them. For example:

(1) The validity of an achievement test is the extent to which the content of the test represents a balanced and adequate sampling of the outcomes (knowledge, skills, etc.) of the course or instructional program it is intended to cover (content, face, or curricular validity.) It is best evidenced by a comparison of the test content with courses of study, instructional materials and statements of instructional goals, and by critical analysis of the processes required in responding to the items.

(2) The validity of an aptitude, prognostic, or readiness test is the extent to which it accurately indicates future learning success in the area for which it is used as a predictor (predictive validity). It is evidenced by correlations between test scores and measures of later success.

(3) The validity of a personality test is the extent to which the test yields an accurate description of an individual's personality organization (status validity). It may be evidenced by agreement between test results and other types of evaluation, such as ratings or clinical classification, but only to the extent that such criteria are themselves valid.

The traditional definition of validity as "the extent to which a test measures what it is supposed to measure," seems less satisfactory than the above, since it fails to emphasize that the validity of a test is always specific to the purposes for which the test is used, and that different kinds of evidence are appropriate for appraising the validity of various types of tests.

Validity of a test item refers to the discriminating power of the item-its ability to distinguish between persons having much and those having little of some characteristic.

APPENDIX

The examples included in this section are portions of scales developed to measure some of the objectives in specific projects. Although the format can be copied, the objectives to be measured must dictate the content of new scales.

GEOMETRY ATTITUDES SCALE
(method)

The following items have been prepared to permit you to indicate how you feel about your geometry class. Your answer is correct if it expresses your true opinion. PLEASE ANSWER EVERY ITEM. In each case, draw a circle around the letter which represents your own ideas as follows:

SA	if you strongly agree with the statement
A	if you agree but not strongly so
U	if you are undecided or neutral
D	if you disagree but not strongly so
SD	if you strongly disagree with the statement

1. I like the way geometry was taught this semester....... SA A U D SD

2. I would have liked to ask more questions during this semester.. SA A U D SD

3. I knew how I was doing in geometry all semester........ SA A U D SD

4. The grading has been fair this semester................ SA A U D SD

5. Students really paid attention to the teacher in this class.. SA A U D SD

6. There was a lot of class time wasted this semester..... SA A U D SD

7. I wasn't able to keep up with the other students this semester... SA A U D SD

8. It took too long to get my test papers back in this class.. SA A U D SD

9. We covered the subject too fast this semester.......... SA A U D SD

10. I had plenty of opportunities to work on my own this semester... SA A U D SD

11. I believe that too much written work was required...... SA A U D SD

12. I worked more in geometry than in other classes this semester.. SA A U D SD

13. I think more use of teaching aids (charts and illustrations) should have been made.................. SA A U D SD

14. Too much outside work was required in geometry......... SA A U D SD

15. There was too much emphasis on things that weren't important this semester.............................. SA A U D SD

16. It was too easy for the slackers to get by this semester... SA A U D SD

ENGLISH ATTITUDE SCALE
(teacher)

The following items have been prepared to permit you to indicate how you feel about your English teacher. Your answer is correct if it expresses your true opinion. PLEASE ANSWER EVERY ITEM. In each case, draw a circle around the letter which represents your own ideas as follows:

SA if you strongly agree with the statement
A if you agree but not strongly so
U if you are undecided or neutral
D if you disagree but not strongly so
SD if you strongly disagree with the statement

1. Miss Montgomery knows a lot about English.............. SA A U D SD
2. I feel that Miss Montgomery is interested in students.. SA A U D SD
3. Miss Montgomery talks over the heads of most of the class.. SA A U D SD
4. Miss Montgomery is well liked by everyone............. SA A U D SD
5. Miss Montgomery seldom got impatient this semester..... SA A U D SD
6. Miss Montgomery really encourages students to think.... SA A U D SD
7. Miss Montgomery was always pleasant this semester...... SA A U D SD
8. Miss Montgomery explains each lesson thoroughly........ SA A U D SD
9. I think that Miss Montgomery is an excellent English teacher.. SA A U D SD
10. Miss Montgomery has a pleasant voice.................. SA A U D SD
11. Miss Montgomery made me feel that I learned a lot in this class.. SA A U D SD
12. Miss Montgomery seemed to enjoy teaching this semester. SA A U D SD
13. Miss Montgomery has many different ways of explaining a difficult point.................................. SA A U D SD
14. Students really respect Miss Montgomery............... SA A U D SD
15. Miss Montgomery was sometimes pretty vague in explaining things.................................... SA A U D SD
16. Miss Montgomery used examples that were meaningful to me.. SA A U D SD

TEACHER CHARACTERISTICS

The following items contain pairs of words describing people. The words in each pair have nearly opposite meanings. Separating the pairs of words are five spaces which can be used to indicate the degree to which each word describes a person. You are to place an X in the space where you think Mr. Wells (Miss Montgomery) would be best described. For example,

tall _____ __X__ _____ _____ _____ short

the X in this space would mean that he is taller than average; or the X in the following space would mean that he is very short:

tall _____ _____ _____ _____ __X__ short

PLEASE ANSWER EVERY ITEM

talkative _____ _____ _____ _____ _____ quiet

dull _____ _____ _____ _____ _____ interesting

friendly _____ _____ _____ _____ _____ cool

excitable _____ _____ _____ _____ _____ calm

polished _____ _____ _____ _____ _____ awkward

solemn _____ _____ _____ _____ _____ cheerful

adaptable _____ _____ _____ _____ _____ rigid

nervous _____ _____ _____ _____ _____ relaxed

sociable _____ _____ _____ _____ _____ withdrawn

disorganized _____ _____ _____ _____ _____ organized

quick _____ _____ _____ _____ _____ slow

critical _____ _____ _____ _____ _____ tolerant

patient _____ _____ _____ _____ _____ impatient

lax _____ _____ _____ _____ _____ demanding

humorous _____ _____ _____ _____ _____ dry

harsh _____ _____ _____ _____ _____ gentle

shy _____ _____ _____ _____ _____ bold

strict _____ _____ _____ _____ _____ easy going

ACTIVITIES PARTICIPATION SCALE

A teacher's duties include out-of-class activities. Some activities are carried on with students and others involve representing the school at meetings. Several of these activities are mentioned on this page. You are to think how well Mr. Wells (Miss Montgomery), your geometry teacher (your English teacher), would fit into each of these activities. Draw a line under the phrase below each activity which best describes how you feel about him (her).

HOW WOULD YOU LIKE TO HAVE MR. WELLS (MISS MONTGOMERY)

1. as your class sponsor?

   Very well Fairly well Neutral Not very well Not at all

2. eat dinner with your family?

   Very well Fairly well Neutral Not very well Not at all

3. represent your school at a state convention?

   Very well Fairly well Neutral Not very well Not at all

4. sponsor a trip to Omaha?

   Very well Fairly well Neutral Not very well Not at all

5. be faculty advisor to your school club?

   Very well Fairly well Neutral Not very well Not at all

6. chaperone a school party?

   Very well Fairly well Neutral Not very well Not at all

7. give a talk to parents at a Parent Teacher Association meeting?

   Very well Fairly well Neutral Not very well Not at all

TV CHECKLIST

Name ______________________________ Date ____________________
Boy ____________) Check one
Girl ____________) Age ____________ Grade ________
School ____________________________ Teacher __________________

Directions: Do you watch any of the programs in the following list? Mark those programs with a:

1. if you like the program so much that you usually watch it
2. if you like to watch the program sometimes
3. if you watch the program when you have nothing else to do.
4. if you never watch it.

Write in the names of other programs that you think should be included in this list, and mark them also with a 1, 2, 3, or 4 to show how much you like them.

MOVIES

__Million Dollar Movie
__Late Show
__Channel 4 Movies
__Channel 5 Movies
__Channel 7 Movies
__Channel 9 Movies
__Channel 11 Movies

____
____

WESTERN

__Branded
__Gunsmoke
__The Loner
__Lawman
__Cheyenne
__The Deputy
__Hank
__Big Valley
__The Virginian
__Rawhide
__Bonanza

____
____

MUSICALS--VARIETY

__Hullabaloo
__Shindig
__King Family
__Steve Lawrence Show
__Andy Williams Show
__The Tonight Show
__Hollywood Palace
__Walt Disney's Wonderful World of Color
__Lawrence Welk Show
__The Ed Sullivan Show
__Clay Cole's Diskotek
__Soupy Sales
__Jimmy Dean Show
__Lloyd Thaxton Show
__The Danny Kaye Show
__The Red Skelton Hour
__Danny Thomas Special

____
____

DRAMA

__Perry Mason
__Twilight Zone
__Peyton Place
__Hawaiian Eye
__Bob Hope Presents
__Colt 45
__Dr. Kildare
__Combat
__Profiles in Courage
__Ben Casey
__Run For Your Life
__Breaking Point
__A Man Called Shenandoah
__Checkmate
__Naked City
__Slattery's People
__The Man from U.N.C.L.E.
__Arrest and Trial
__Convoy
__The Long Hot Summer
__Richard Boone Show

____
____

U.S. GOVERNMENT PRINTING OFFICE : 1966 O—236-150